Organisational aspects of implementing ML based data editing in statistical production

modernstats
by HLG - MOS

Acknowledgement

Machine Learning (ML) holds significant potential for enhancing efficiency by complementing or replacing traditional methods, as well as improving quality in ways that are challenging for traditional methods to achieve. However, numerous barriers hindering statistical organisations from implementing ML methods for editing are non-methodological. The objective of this document is to identify these barriers and offer guidance on avoidance or overcoming them. With this document, we aim to encompass discussions and advice on addressing the identified issues, along with a compiled set of use cases gathered during the task team work.

We thank the following team members for their generous dedication of time and for contributing their valuable input:

- Claire Clarke (chair) and Jenny Pocknee –Australian Bureau of Statistics
- Wesley Yung, Jean Le Moullec and Stan Hatko –Statistics Canada
- Riita Piela –Statistics Finland
- Steffen Moritz – German Federal Statistical Office
- João Poças –Statistics Portugal
- Sandra Barragán, David Salgado and Elena Rosa-Pérez –Statistics Spain
- Jens Malmros –Statistics Sweden
- Daniel Kilchmann – Swiss Federal Statistical Office
- Olivier Sirello and Bilyana Bogdanova – BIS
- Amilina Kipkeeva – UNECE

Table of Contents

1. Introduction

Data cleaning and editing are essential components of ensuring the quality of official statistics. However, finding and correcting errors in datasets can be a lengthy and time-consuming process. The increasing size of modern datasets makes manual interventions increasingly infeasible, and new types of data may not be well served by traditional methods. Machine learning methods have strong potential to provide solutions to these challenges.

Editing and imputation were identified as some of the most obvious use cases for machine learning in the first High-Level Group for the Modernisation of Official Statistics (HLG-MOS) project on machine learning, conducted in 2019 and 2020. The report from this project concluded that editing and imputation were valid use cases for machine learning in the production of official statistics[1]. However, since then, agencies have been slow to adopt machine learning methods for editing. Rather than discussing or proposing technical approaches to editing using machine learning, the Applying Data Science and Modern Methods (ADSaMM) Data Editing task team decided that it would be more valuable to examine some of the blockers preventing the adoption of these methods and suggest some guidelines for overcoming them. We decided to pursue this by gathering use cases from official statistics agencies to understand what the biggest difficulties were and how they had been overcome in each agency.

The process followed by the task team was as follows:

1. We developed a template for gathering use cases. This was initially framed around the steps in the journey from experimentation to development for machine learning methods described in Chapter 5 of Machine Learning for Official Statistics[2]. Some adjustments were made to this initial version after gathering a couple of examples and determining which elements the task team found most useful. The team also eventually decided to incorporate a short technical description of the methods used in each use case, as this was deemed of considerable interest. The template is included in Appendix 2 for reference.
2. We identified potential use cases and reached out to the agencies involved to fill in the template. Early members of the team provided a small number of use cases that were used as examples to assist subsequent use case development. The most fruitful source of intelligence for identifying potential use cases was the agenda of the UNECE Machine Learning for Official Statistics 2023 Workshop[3]. In most cases, agencies that provided a use case also gave a short presentation to the team about their use case and provided a staff member to join the team. In the end, there were seven use cases overall.
3. We assessed the use cases to extract key themes and identify areas where there were blockers to the implementation of machine learning for editing. We utilised the use cases to craft brief descriptions of these key issues and provide

[1] https://statswiki.unece.org/display/ML/Machine+Learning+Project+Report
[2] https://unece.org/sites/default/files/2022-01/ECECESSTAT20216.pdf
[3] https://unece.org/statistics/events/ML2023

guidelines for overcoming them. Then, we edited these guidelines along with a select set of use cases (those which agencies agreed could be included) to create a coherent document.

This document is the outcome of that work. Sections 2.1.-2.6. contain reflections on each of the key issues we identified. Appendix 1 contains information on the implementation of ML Operations (MLOps). Appendix 2, as noted above, contains the template used for constructing the use cases, while Appendix 3 contains the complete set of use cases gathered as part of the task team work. We hope you find this information useful.

The chair would like to thank all members of the Data Editing task team for their contributions to this work, and to extend special thanks to all the agencies that supplied use cases.

2. Key Themes

Based on the use cases assembled, the task team identified six influential factors that contribute to the adoption of machine learning editing methods. These were:

- The driver of the problem being addressed by a machine learning solution
- The lack of labelled data or other suitable training data
- The relationship between business area, methodologists/data science staff and IT specialists
- The need for input and feedback from subject matter specialists
- Domain specific knowledge and the black box nature of machine learning methods
- IT issues and Machine Learning Operations and machine learning platform

Each of these will be addressed separately in the following short sections.

2.1. Driver of problem

An influential factor in whether or not a machine learning solution will be accepted and progressed into production is the driver of the problem. This is especially the case for editing, where business areas may have confidence in human-led quality assurance, and less confidence in or understanding of the methods underpinning machine learning. Considerable motivation is required to change approaches even where a change may lead to substantial efficiencies. **Business areas may be less open to proposals for change that are not driven by their needs or take an ”if it isn’t broken don’t fix it” attitude when they are satisfied with the quality of their current methods.**

For these reasons, it may be necessary to look for the right kinds of opportunities to introduce machine learning methods for editing or to reframe a problem so that it presents as the right kind of opportunity. The two most commonly represented in the use cases are:

1. **Acquisition of new data:** In this scenario, an organisation receives, or is anticipating the reception of, a new dataset for which traditional approaches to editing will be less than optimal. This is most likely (for example, in the Australian Bureau of Statistics (ABS) and Statistics Portugal use cases (Appendix 3, use cases 1 and 3) because the dataset is very large. Editing methods that employ human intervention become infeasible at the scale of some larger administrative datasets that have high frequency and/or high volumes of data. It is also possible that new datasets may contain types of data

that cannot be edited via traditional methods, such as text data. The Bank for International Settlements (BIS) use case (Appendix 3, use case 6) falls into this scenario - although not a new situation, there were certain kinds of time series that could not be quality assured using traditional methods. These new situations present substantial opportunities to demonstrate the capabilities and benefits of machine learning methods, as these are designed by their nature to handle large amounts of data and non-standard data in ways that traditional methods are not. In other words, in these scenarios, machine learning methods present a very natural solution to the problem at hand.

2. **Improvement to current methods:** Situations where there is clear evidence that current methods have some deficiency, whether that be in accuracy, speed, or coverage, also present opportunities to introduce machine learning methods. Statistics Canada, Swiss Federal Statistics Office and Statistics Sweden use cases (Appendix 3 use cases 2, 4 and 5), fall into this kind of scenario. In these situations, there is less of a case for machine learning to provide a natural solution to the problem at hand. It may therefore be necessary to compare the machine learning method to the old approach and/or to another non-machine learning method in order to convince business areas that the additional complexity of introducing machine learning is worthwhile.

One further scenario, represented by the Statistics Spain use case (Appendix 3 use case 7), is the opportunity to develop new products or services. In this situation, it may also be less clear that machine learning methods are a natural solution to the problem being addressed (although this depends on the nature of the problem). Again, it may be necessary to compare them to traditional methods to show that they offer superior performance.

2.2. Lack of training/labelled data

In the realm of data editing, one central pillar that ensures the accuracy, relevance, and integrity of automated edits is the availability of abundant, high-quality training data. However, the reality often presents a quite different picture with issues stemming from insufficient and unlabelled data. These issues often snowball into formidable challenges affecting the development, efficiency, and effectiveness of (machine learning) models applied in data editing.

1. **Lack of (high quality) labels:** Typical problems here are "Absence of labels", "Low quality/biased labels" and "Late availability of labels". Since ML-based solutions for (automated) data editing usually need labels either for model building/training or evaluation, the lack of high-quality labels can have a huge impact. The problem affects ML solutions for detecting errors as well as ML solutions for correcting errors. This is because, bias in the data/labels will often be propagated to the model. In the following, each problem is described in detail with potential mitigation measures.

- Absence of labels

  In supervised learning, the absence of labels hinders the development and evaluation of (machine learning) models. When there are no (or not enough) labels for the target variable, a sufficient model cannot be built. A lack of labels makes it essentially impossible to learn the connection and patterns between predictors and the target variable. Relying more on unsupervised learning often does not necessarily solve the problem. In unsupervised learning, the absence of a ground truth (labels) can turn the identification of automated editing candidates into a complex puzzle. Distinguishing errors from non-errors can be an uphill task, considering not all extreme values amount to errors, and vice versa, not all errors manifest as extreme values. Examples where there are not enough labels are usually missing data problems. Some labels are present (the complete data), which are used for building the imputation model. But the missing data itself is quite often not recoverable (forever unknown), which complicates the evaluation of the imputation results.

  Possible Mitigations: Make an honest approach to derive quality labels, e.g., with the expertise of a human reviewer. Combine unsupervised methods with specialised knowledge, use human-in-the-loop approaches for critical / influential edits, or use simulation studies (overimputation).

- Low quality /biased labels

  In addition to missing labels being a problem, another very common problem is biased labels. That is, the labels themselves are there, but their manifestations can be incorrect or there may be no real consensus about their values. This can be seen when different people labelling the data come to different conclusions. Ultimately, this means models built on these labels will be biased, which affects the subsequent data editing actions.

  Possible Mitigations: Analyse reviewer consensus, find solutions for reviewer consensus, put more effort into label quality, use methods for uncertainty quantification or use human-in-the-loop approaches for some edits.

- Delayed availability of labels

  Delayed availability of labels also affects development and evaluation of models and editing solutions. A key difference to the mere complete/partial absence of labels is that the labels do become available at some point in time. But, usually too late to include them in the data editing process for the currently ongoing statistical production process.

In comparison to the absence of labels, late incoming labels at least enable some kind of ex-post evaluation, which would not be possible otherwise (e.g., for missing data).

Possible Mitigations: See also all suggestions for absence of labels, try to speed up label availability, or work with partial label deliveries.

2. Lack (quality, amount) of training data

- Not enough data

  Not having enough data can affect the performance of the (machine learning) model used for data editing. Very limited data leads to a whole list of problems. For example, overfitting might become a problem, because the model learns (or focuses too much on) the noise and outliers of the limited data instead of generalising. Furthermore, some potentially predictive feature manifestations / variations might not appear in the limited data, preventing the model from using them effectively. Also, model/parameter selection becomes difficult, since there is only a limited number of test/train evaluation combinations. Overall, this causes the underlying models to fall short of the required level of robustness and accuracy.

  Possible Mitigations: Obtain more data, leverage known approaches from the ML literature such as data augmentation, bootstrapping or some form of transfer learning,

- Non-representative data

  If the data comes from a non-probability sample, certain groups in the data could be overrepresented. This may lead to bias because the dataset does not represent the entire population. The model trained on the data may not generalise well to the broader population - leading to skewed predictions. Unfavourably, the evaluation metrics may also be misleading since they would be computed on the same biased data and not on the general population. Overall, the models could be skewed, exhibiting partiality towards particular trends, patterns, or classifications as a result.

  Possible Mitigations: Obtain more data, ensure that the training data does not suffer from selection bias, use statistical methods to mitigate sampling bias or control the selection of training data.

- Delayed availability of data

Delayed availability (of some) of the data basically comes with the same issues as "Not enough data" and "Non-representative data".

Possible Mitigations: See also all suggestions for "Not enough data" and "Non-representative data", attempt to speed up data availability or work with partial data deliveries.

1.3. Relationship among business areas, methodology, data science team(s) and IT specialists

For many years, National Statistical Organisations (NSOs) have been applying statistical methods to produce high quality outputs from data typically obtained from surveys or administrative sources. With the expansion of data science tools, in particular machine learning, NSOs are exploring ways to integrate these tools into their production processes. **The challenge has been that the dynamic between subject matter experts (the "business"), methodologists and IT specialists is already well established in the organisation, and the data science group has had to integrate themselves into this dynamic.** As illustrated in the use cases in Appendix 3, this has been successfully done in some situations and less so in others. This section will provide some themes pulled out from the use cases and potential best practices.

A common thread in the use cases was that the business areas usually came to the data science areas looking for a solution to a particular problem. While this is encouraging, it can lead to a relationship where the data science area is seen almost as a 'consultant' who has been hired for a particular task. Methodology groups have been particularly successful as they are known as an area which can solve many different problems related to statistical methods. If a data science area can gain a broader reputation, it will help with having them consulted on more varied problems. In addition, if a data science area can become familiar with the business area and other problems that they are facing, then they may be able to offer solutions to those problems.

All use cases recognise that close cooperation between the data science group and the business area is essential. Several use cases (Australia, Portugal, and Spain) highlighted that not only does the business area need to understand what the data science area is putting into place, but the data science area needs to understand requirements of the business area.

In the use cases, the relationship between methodology and the data science areas is not always clear. Most use cases mention the importance of collaboration between methodology and the data science area (if one exists) but did not elaborate on it. At Statistics Canada, the data science area is housed in the same organisational unit as the methodology group to foster collaboration. The methodology group is well integrated into the statistical programmes and steps are underway to leverage this to further integrate the data science group into these programmes. In addition, this

arrangement is helpful in sharing knowledge on both sides and, more importantly, identifying potential barriers to fully integrating data science tools into statistical programmes.

This arrangement has also brought up some interesting discussions on the future relationship between methodology and data science. In recent years, new methodology recruits often join with some competencies in data science. If this trend continues, how will the roles and responsibilities evolve going forward? One possible scenario is that "citizen data scientists" will be more common in both methodology and subject matter areas and that a small data science division consisting of more research-oriented data scientists will be established. This scenario would be similar to what probably happened many years ago as statistical sampling techniques or complex statistical analyses were adopted. However, both of those examples took multiple years to occur.

Similar to the relationship between methodology and data science areas, the one between IT and data science has also been a challenge. The major challenge has been the concerns around IT security and the ability to provide the necessary IT infrastructure for the new data science applications such as computing power, data storage. There will obviously be a "feeling out" stage where IT and data science will have to learn about each other and to define roles and responsibilities, but the earlier that this is achieved the better for the organisation. The advent of ML Operations (MLOps) has brought a new dimension to the traditional Development and Operations (DevOps) framework. Often misunderstood as competing approaches, DevOps and MLOps are, in fact, deeply interconnected, each playing a pivotal role in the lifecycle of software and ML development. DevOps and MLOps share foundational principles of automation, iterative processes, and a collaborative ethos. The Continuous Integration/Continus Deployment (CI/CD) pipelines, central to MLOps, are predominantly an extension of DevOps practices, underscoring the interplay between the two. When aligning MLOps with DevOps, it is imperative to delve into the intricacies of MLOps. This understanding is pivotal in recognising how MLOps does not just coexist with DevOps but actively intertwines with it, enhancing and extending its capabilities. MLOps aims for the automation of processes and champions transparency and reproducibility, aligning closely with the core objectives of DevOps.

1.4. Input and feedback from subject matter experts

Relations among different profiles in the statistical offices are not always easy to manage and getting to a full understanding that leads to fruitful results can be difficult. However, these multidisciplinary teams are the key to success.

Subject matter experts have accumulated great knowledge in the particular statistical areas for which they are responsible for. Their competencies and skills have been developed through years of training and experience learned while working, even accumulated from former colleagues in the same subject matter. In relation to subject matter experts, the challenge is the lack of time due to the production process. They

are focused on the needs of production that sometimes require urgent interventions, so it is difficult for these experts to be engaged in innovation projects. At the same time, they have a great amount of knowledge about the real needs of production, and they know the behaviour of the data better than anyone so they can be helpful with interpretability of intermediate results which gives feedback to improve the methodology. Subject matter experts are vital also in the first steps of the machine learning methods with their description of the manual procedures to be transformed into regressors containing essential information for the model. Another challenge arises from the steep learning curve of new methods, making the project appear too difficult to confront for the subject matter experts who are not familiar with the new methods.

From the point of view of the methodology units, it is important to understand the problems and needs of the business areas but even more important to be able to develop standard solutions that solve not only the problem at hand but similar (of the same nature) problems that could appear in other business areas. Then, a possible fruitful collaboration is not one-to-one but the hub and spoke model to build teams where the methodology unit is in the root and the business areas are in the nodes. Then, the methodologists can understand not only the initial problem but others that are of similar nature.

Potential mitigation measures include:

- Incorporate subject matter experts right from the project's inception. Ensure that they are not just participants but actively recognized as integral contributors to the project.
- Engage the subject matter experts at the early stage is also important to learn their real needs and incorporate them in the design of solutions.
- Explain the methodology to the subject matter experts and give them enough training in order to feel comfortable with the new process that they will have to run.
- Transmit to the subject matter experts that these new projects are an opportunity for them to improve and to save time and encourage them to see the time spent for the project as an investment for the future. Starting from recent new methods incorporated to the pipeline process as cases of success, set how the result of this new project will be in the production process and which are the advantages of that.
- Work with groups of people within the structure of the hub and spoke model. (See “The Use of Data Science in a National Statistical Office”, Erman et al (2022))

1.5. Requirements for data science expertise and black box issues

One key aspect related to the implementation of ML pipelines into a production environment relates to the organisation readiness, including the human resources

(e.g., availability of expert, trained staff in data science and machine learning). This is critical not only to grasp all the benefits from using ML for statistical production, but also to prevent black-box challenges, that is, the use of obscure ML algorithms.

To start with, ML methods often require strong expertise in data science. Several use cases in Appendix 3 indicate the specific need for knowledge, training, and recruitment in order to keep up with ML advances in a rapidly changing environment. For instance, the Australian Bureau of Statistics (ABS) mentions the effort to provide staff with knowledge about data science activities and the use of machine learning methods. Statistics Portugal (INE) also reports a very similar requirement, with its management encouraging training courses in data science, both to empower employees with new knowledge and to deploy machine learning methods into their daily work activities. Conversely, only few organisations (Statistics Sweden) admit having sufficient knowledge in-house to build and maintain ML-based applications for official statistics.

The transparency of the ML methods chosen is also key to prevent black box issues. This is crucial in order to mitigate operational and reputational risk for the organisation in case ML pipelines generate unexpected results which cannot be explained. To this respect, some organisations promote synergies between data scientists, IT and business areas in order to conduct an in-depth evaluation and test phases of the methods chosen as well as to jointly define validation, consistency and coherence analysis steps (e.g., Statistics Portugal and Statistics Spain). Code sharing is another approach to mitigate the black-box risk followed by several organisations such as the Bank for International Settlements. It aims to eventually foster discussions among experts on the best methods to follow and avoid the use of highly uncertain, complex algorithms. Organisations may also consider the possibility to disclose the full decision-matrix behind the usage of machine-learning techniques, including the rationale behind the selection of specific parameters. Finally, black box machine learning models and their potential model failures can be mitigated by setting up rigorous uncertainty sets (e.g., conformal prediction) for the predictions of the models used in production.

1.6. IT issues (ML infrastructure)

As mentioned, IT infrastructure, systems and processes are fundamental to harness data science, machine learning and compute capabilities. While the adoption of emerging data science technologies offers potential opportunities, such as meeting the computation demands of big data, there are also challenges. These challenges are relevant for innovations generally but appear particularly so for machine learning data editing projects. These challenges can also depend on where an NSO is on their IT / data science modernisation and machine learning journeys. This section outlines some of the key issues and potential solutions. For more background on challenges to machine learning projects, please refer to the “Building an ML system in Statistical

Organisations"[4] report from the Office of National Statistics Office (ONS)-UNECE ML Group 2022.

Innovation projects in general require IT systems that support the research and development process. Innovation is more likely to be successful if an organisation has streamlined R&D environments / processes that support the innovation cycle, for example, environments that enable data to be brought together with emerging tools and software in a safe way. Research environments may have less functionality than production systems; so later stages of the innovation cycle might require additional assessment be undertaken, for example, model hyperparameters, and compute performance using full-scale data. It is important to allow for these steps.

Another aspect of the innovation cycle is the importance of streamlined governance processes, such as resourcing different stages of the cycle and go/no go decision-making. It is particularly important that "production owners" for the methods, IT, data science and statistical subject matter have been identified and agreed upon early in the process. For example, Statistics Canada uses formal Service Level Agreements for production roles and responsibilities across production staff, IT staff, data scientists and statisticians.

The innovation may also require integration with, and modifications to, the production environment. Productionisation takes effort and resourcing to test, deploy and integrate the model / components into the proof of concept and production systems; including refactoring code to reduce tech debt, automation (e.g., iterative model updating), memory usage, and I/O optimisation. This integration with production may also include components such as pre-processing, Quality Assurance/Machine Learning tools (Statistics Canada), related editing/imputation processes and tools (such as manual editing) and incorporating any necessary system changes to standard outputs such as prediction errors (Statistics Spain). Some components or underlying processes may not yet exist in a production system for an organisation, such as R/Python servers or cloud compute capabilities. It is important to start arranging production IT infrastructure early because of the time and resourcing demands on IT teams.

Many NSOs are undertaking IT / data science modernisation programmes, which provides opportunities for innovation and enables the organisation to meet future needs. However, it also places high demands on IT teams as modernisation programmes can be a long multifaceted journey that stretches IT teams' support over new and existing systems through the transition. Innovations beyond these programmes may compete with these resources, and so need to be seen as complementary. The emerging tools and supporting infrastructure need IT staff to build components and provide ongoing support.

[4] https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022

Cloud-based environments provide the potential to manage big data and harness emerging technologies and open-source software. While this can be the catalyst and opportunity for editing and imputation projects, there are some challenges.

For example, different cloud providers offer different services / functionality; what an organisation wants (e.g., MLOps_ may not be easily available. Standard production system components and tools may not be easily incorporated, for example, not all programming languages are natively supported. It takes time and resources to adapt cloud environments to meet the needs of an NSO, to build and incorporate these components and aspects (e.g., security). This means that environments under development may not (yet) have all the services and functionality needed for ML data editing.

Acquiring and developing skillsets is essential. Collaborating with cloud providers can be beneficial, although the potential issue of vendor lock-in should be considered. These environments provide access to open-source programming languages with a wide range of packages that may be available, which is useful for ML projects. While in-built machine learning cloud services may be available, organisations need to consider the needs of an NSO for transparency, explainability and control (for more refer to the HLG-MOS Project "Cloud for Official Statistics" (2023))[5].

Support for programming languages: Each programming language used by an organisation requires a support team, so NSOs may select a set of languages to support. Every programming language has strengths and limitations. For example, SAS is a trusted and well-supported programming language widely used for official statistics. Open-source programming languages such as R and Python offer a wide range of pre-built packages that are useful for statistics (including those developed by NSOs) and are particularly useful and flexible in the ML space. Being open-source, there is no guarantee over robustness and support for packages. Nevertheless, many R and Python packages do have committed support teams. They also require more effort on the part of the organisation, for example, version management.

Not all functionalities can be met by pre-built software / packages, so some components may be developed or modified in-house. These custom solutions take additional effort to build and maintain. This may especially be the case to adapt emerging approaches such as machine learning, to meet the needs of an NSO such as applying for a statistical product and providing greater control or explainability.

Vendor lock-in: Historical decisions about the IT environment may make it harder to incorporate emerging technologies particularly when adapting or transitioning away from legacy systems. For data editing projects, this could apply, for example, to the introduction of open-source programming languages (and the supporting

[5] The current version is available here, the document to be updated: https://unece.org/sites/default/files/2023-12/HLG-MOS2023%20Cloud%20for%20Official%20Statistics_DRAFT.pdf

infrastructure and processes), ML infrastructure (refer to the Appendix 1), and cloud environments/tools. For example, the "Cloud for Official Statistics" project noted the importance of having an exit strategy from the start when procuring IT solutions, so that costs are understood (such as egressing data), and that time and resources are already allocated to transitioning at a later stage. For cloud solutions, what is possible in terms of an exit strategy depends on the type of cloud approach that the organisation has adopted. The project team also noted that vendor-agnostic and open-source culture also make it easier to acquire skilled staff. Cloud-related skills are in high demand and take time to build, so it can be useful to work closely with vendors to develop systems. However, one needs to be mindful of the potential for lock-in if vendor-specific systems are embedded and skills are developed in the organisation.

3. Conclusion

In previous sections, we have discussed a range of non-technical issues that can present barriers to introducing machine learning methods for editing. Some of these issues may also arise when trying to use ML methods for other parts of the statistical production process. For all these issues, it is clear that careful planning of any machine learning project is necessary for success. This might include factoring in costs to acquire and label training data, planning to bring in different teams with different expertise at the right points in the project, scheduling project milestones to facilitate the involvement of busy subject matter experts, or including time and resources to educate future users. Building good relationships and having clear lines of responsibility between methodology, data science, business and IT teams is vital, as is making sure that appropriate IT environments and resources are available to support the demands of an ML project. Different approaches may be needed to carry out a proof of concept compared to introducing methods into production, but the latter should not be ignored when planning the former in order to facilitate a smoother introduction into production later.

The implementation of ML Operations (MLOps) in statistical offices represents a pivotal transition from traditional ways of implementing methods to more advanced and automated processes. MLOps also provides a structured framework for embedding Responsible AI principles. These principles ensure ethical, transparent, and accountable use of AI and machine learning, covering aspects such as fairness to prevent biases in models, accountability in development and deployment, transparency in decision-making processes, adherence to ethical standards in data usage and protection of sensitive information. MLOps in statistical offices is not just a technological upgrade but a comprehensive strategy towards more responsible, efficient, and advanced data processing and analysis. Along with responsible AI and other MLOps principles, it covers IT infrastructure, tools, processes, and roles. This transition is essential for statistical offices to remain relevant and effective in a data-driven era, guaranteeing the provision of accurate, reliable, and insightful statistics.

Because of this we include some reflections on what is needed to set up MLOps in the Appendix 1.

Appendix 1: Implementing MLOps in a National Statistical Office

Objective: To establish an MLOps that ensures the seamless integration, deployment, monitoring, and maintenance of ML models while adhering to the principles of accuracy, privacy, transparency, and reproducibility required.

Factors to consider:

1. Data collection and management:
    - Ensure data anonymisation and encryption to maintain privacy (usually operated in the cloud).
    - Use version control for datasets to track changes and updates (needed for reproducibility).
    - Add data quality assessment steps to ensure data's accuracy.
2. Model development and validation:
    - Set up a development environment/platform with tools like Jupyter notebooks or RStudio.
    - Use version control (e.g., Git) for model code to ensure reproducibility.
    - Implement a model validation framework to ensure models meet accuracy and reliability standards before deployment.
    - Highlight the importance of cross-functional collaboration between different roles (data scientists, ML engineer and domain experts).
    - Stress the importance of a standardised model development framework to ensure the consistency and ease of validation.
3. Automated testing:
    - Develop automated testing pipelines to validate data processing scripts and ML models.
    - Include tests for data quality, model accuracy, and performance benchmarks.
    - Set the baselines for model performance metrics to compare the outcomes of automated tests.
4. Continuous Integration and Continuous Deployment (CI/CD):
    - Implement CI/CD pipelines using tools like Jenkins, Azure DevOps, or GitHub Actions.
    - Ensure automated testing is integrated into the CI/CD pipeline.
5. Model monitoring and maintenance:
    - Monitor model performance in real-time using tools like MLflow or Prometheus.
    - Set up alerts for any significant deviations in model performance.
    - Implement a retraining pipeline for models to ensure they remain accurate as new data becomes available.
    - Specify the metrics to be monitored for model performance.

6. Documentation and compliance:
    - Maintain comprehensive documentation for all data processing and ML workflows and models.
    - Ensure compliance with national and international standards for data privacy, security, and ethics.
    - Implement audit trails for all data and model operations.
    - Ensure the versioning of models, data, and code (enabling reproducibility).
7. Stakeholder communication:
    - Develop dashboards using tools like PowerBI or Tableau to communicate model results and insights to stakeholders.
    - Ensure transparency in model decisions and provide explanations where needed.

The machine learning platform provides a scalable environment that supports diverse stages of ML model development, deployment, and maintenance. Key features include:

- Data processing and storage: systems for handling large volumes of diverse data, with high-performance computing capabilities.
- Development environments: integrated tools like Jupyter Notebooks and RStudio, facilitating collaborative development and experimentation.
- Model training and testing: advanced GPU-accelerated hardware for efficient model training and testing.
- Deployment and monitoring: infrastructure to deploy models in production and tools to monitor their performance continuously.
- Security and compliance: strong security protocols and compliance mechanisms to protect sensitive data and adhere to regulatory standards.

Technologies:

- Cloud platforms: AWS, Azure, Google Cloud for scalable, on-demand compute resources.
- Version control: Git for code, DVC (Data Version Control) for data management.
- CI/CD tools: Jenkins, Azure DevOps, GitHub Actions for continuous integration and deployment.
- Monitoring tools: MLflow, Prometheus for real-time performance monitoring.

MLOps Role responsibilities (examples):

- Data Scientists: focus on model development, data analysis, and algorithm selection. Responsible for initial data pre-processing and exploratory data analysis.

- ML Engineers: specialise in refining ML models for production, optimising algorithms and implementing efficient data pipelines.
- DevOps Engineers (can also be L Engineers): manage the CI/CD pipeline, ensure infrastructure health and oversee the deployment and scaling of ML models.
- Security Specialists: ensure the security of the ML platform and compliance with data privacy and protection standards.
- Domain Experts (stakeholders): provide domain-specific insights and validate the relevance and applicability of ML models to organisational objectives.

Appendix 2: Use Case Template

With the explosion of data now available, modern methods such as machine learning have gained significant traction in everyday life by companies such as Google, Amazon, and Microsoft amongst many others. However, the same can not necessarily be said when it comes to National Statistical Organisations (NSOs) where the uptake has been less than in the private sector. This template is to gather some insight on why the uptake by NSOs has been slower than in the private sector and how it could be accelerated.

Guiding questions:

1. Throughout the journey from experiment to production, what has been stopping your organisation from applying data science and modern methods in data editing?
2. Focusing on the organisational aspects along the productionisation process, how did the project manager overcome those obstacles (e.g., how to explain the processes and methods of data editing to stakeholders within and beyond the organisation and proof that it is a good value for money)? Please identify the problems and stakeholders involved (e.g., the dynamics between senior management, research team, business area, IT team, etc.) and the actions taken to resolve the issues and improve multi-level engagement throughout the different stages listed in the use case template.
3. To help contextualise the use cases, please discuss the pros and cons of applying these modern methods in data editing (in terms of accuracy, explainability, transparency, cost effectiveness, or other metrics) that help get buy-in from stakeholders.
4. What are the lessons learned and best practices that would be useful for other NSOs?

	Title of the use case and the name of the organisation (Please provide a title of the use case with enough information so that readers can understand the context.)
Project overview	Please provide an overview of the project and describe its strategic importance to the work of the organisation, including the statistical programme in question, the business needs of the project, whether the proposed method is replacing an existing method or is a new application and why a modern method is being considered.
Organisational readiness	When completing this section, please keep in mind the readiness of the organisation related to aspects such as the IT infrastructure, the capacity of the organisation in terms of knowledge of the 'tools' required to set up the method and to also maintain it, as well as the openness of the organisation to

	adopt modern methods
Understand business needs (Who needs what)	Please provide information on the context around motivation of the project. Include information such as the business need, who asked/sponsored/paid for the project and enough information for readers to understand what the real business need is so that they can draw parallels with any projects they may have within their organisation.
Assess Preliminary Feasibility	Please indicate what assessments were made in deciding to investigate the method(s) chosen. These assessments could include considerations such as the appropriateness of the method given the data (continuous vs discrete) or the problem at hand, the availability of required resources (both IT and human resources), and the expected improvements over an existing method (if it exists).
Develop proof of concept	Please provide insight on how the proof of concept was developed and include information such as obstacles and how they were overcome (or not), any 'adjustments' that had to be made to the planned implementation of the method and lessons learned (both positive (keys to success) and negative (blockers)).
Approach/method used	Please provide a detailed description of the approach or method used to develop the proof of concept, model or solution being discussed in the use case. This could include information such as the algorithm used, the data sources and pre-processing techniques, the hyperparameters and training approach, and any other relevant details about the approach or methodology used.
Prepare a Comprehensive Business Case	Please provide insight on what factors influenced the success of the business case, what were the most important components of the business case to achieve acceptance/agreement, were there any obstacles to the preparation of the business case, and how were these overcome.
Deploy the model	Please include the challenges faced in integrating the model into a production system. These could include aspects such as redeveloping the model (to suit production systems and/or data), availability of IT human resources (specialised or not), reluctance to potentially put a production system at risk, availability of specialised IT infrastructure (e.g., ML platform), the need for documentation and training of end users, and availability of funds required.
Results	Please provide information about the outcomes that have occurred if the model has been deployed in production.
Latest status and	Please provide information about the status of the project (e.g.,

next steps	implemented, being programmed into the production system, etc.). If the method is not in production yet, please explain why and share the implementation plan if it exists.
Lessons learned & recommendation	Please consider the following sub-themes: IT infrastructure, IT capacity, organisational knowledge of the proposed method, maintenance of the method once in production and acceptance of the method by business areas.
Reference	Please provide any helpful links and supporting materials.
Contact	Please provide a contact person with an email address.

Appendix 3: Use cases

Use case 1	**ABS: Un-supervised ML for anomaly detection in large and frequent admin data**
Project overview	Anomaly detection of frequent big data – Un-supervised approach to identify anomalies in wage payment administrative data as reported by businesses. The ABS has been investigating unsupervised anomaly detection methods for large and frequent business administrative datasets - wages and jobs as reported by businesses to the tax office, with frequent extracts provided to ABS. Unsupervised methods produce anomaly scores that can be used in combination with significance scores to better-target validation and editing efforts - providing human decision-makers with a short-list of anomalous and significant units, along with contextual information. This forms part of a broader validation and editing approach and is a low-risk way to introduce the benefits of machine learning. The methods were selected based on performance, efficiency and explainability. Unsupervised methods are also useful for identifying unexpected anomalies in new and evolving datasets, where labelled data is limited. This method is being assessed for inclusion in a production system, and if useful may also be considered for other statistical programmes. Other possible future directions may be the automated treatment of less-significant units.
Organisational readiness	The ABS has a long history of innovation. This includes allocating targeted resourcing into key areas of research, methodological developments, and data science / compute capabilities. Ideas are assessed for their potential to concretely improve the delivery of statistical information, such as improving efficiency, quality, capabilities or delivering new statistical insights. This preparedness better-enables the organisation to harness opportunities. This project arose from the use of a big new dataset within a new compute environment - the need to understand and identify anomalies in large, frequent, evolving data. The new environment provided functionality and tools that were not usually available - such as broad access to python/R packages and compute capabilities - which made this work possible for big data. However, the opportunity to undertake this project also came with some limitations. The compute environment was new and still being built, with limited tools, functionality, access, and

	support for early users such as this research team. It was a steep learning curve for all teams involved in this new environment. Before this project started, the methodology area had: - undertaken investigation into potential use of machine learning for data editing more-broadly; and - engaged with other NSOs doing work in this space, including UNECE HLG-MOS and Statistical Data Editing. it was able to leverage these learnings for this project. A key deliverable for this project is to better understand and create documentation for ongoing maintenance of these algorithms, with the aim to gradually build confidence-in-ML and expertise within statistical production areas.
Understand business needs (Who needs what)	As mentioned, the need for this new statistical product in a short timeframe created the opportunity for this project. - This use case involved very large data that needed frequent processing. This provided an opportunity to investigate new compute solutions and machine learning to identify anomalies for big data. Because automated editing rules were already built into the pipeline, with manual validation and editing also undertaken; this project introduced a low-risk complementary approach, aiming to better-target and better-inform the work of the validation / editing team and so provide efficiencies and improved quality. - The team also focussed on building a solution that was useful for the broader organisation (not just a point solution) while also delivering a useful concrete deliverable for the particular use case. - The data itself was large, new, and evolving (with staged onboarding of data providers), thus were still developing our understanding of what 'wrong' looked like. However, this was an opportunity to demonstrate unsupervised methods to identify anomalies, and to help build up our understanding of what 'wrong' looks like. This learning could be used to develop rules or train models to recognise these patterns; however it is anticipated that unsupervised methods would continue to be useful ongoing for identifying unexpected anomalies. This stage of the work was funded through the allocation of a methodology team, with access to the environment and data provided by the statistical production area and IT team. The research team built a good working relationship with the business area that enabled us to understand their needs and put us in a position to be 'on the scene' to provide solutions. A key aim of this project was to build understanding and confidence in the performance of machine learning for this purpose. - To build confidence in machine learning (ML), this work aimed to bring our stakeholders on a journey, starting with a low-risk

	approach to demonstrate that ML can add value / complement the more-traditional and familiar approaches. (As the stakeholders become comfortable that the approach is working appropriately, then later stages may investigate the use of ML to propose edit values.) - A key aspect of this is explainability; It is important to be able to explain how the approach is working and why particular units are being identified, for transparency; to build stakeholder confidence, and to determine whether the approach is working / improve the performance. As mentioned further below, this can be challenging without labelled data. The IT team had been investigating emerging compute environments, so was able to harness the opportunity when it came along. Taking things to the next stage in the productionisation process will depend on future funding decisions.
Assess Preliminary Feasibility	Methods: Building on our existing knowledge of the benefits and issues of ML for anomaly detection, we undertook a small/fast assessment of a number of potential methods and selected a method that suited the nature of the data and production needs - initially applying Local Outlier Factor (LOF). LOF is a density-based clustering method that is relatively efficient; relatively simple to understand and maintain; likely to provide good and robust results with minimal hyper/parameter decision-making and pre-processing. LOF provides a score reflecting how anomalous the unit is, and a subset of anomalous units is sent to the validation team, along with contextual information and visualisations. This shortlist can be further targeted on groups, such as significant contributing units for example. The anomaly scores can be normalised to assist with interpretation (e.g., between 0,1). The initial LOF approach looked only at the current period - that is, in the current period of data, the timeseries variables were created for each unit - and anomalous units were given a higher score if their variable combinations were different other units in the same period. This approach is easy to explain, and easy to maintain as it does not need training data / models to be created. The statistical production area became more familiar and comfortable with the performance and explainability of this approach. Two additional approaches were assessed, both using training data to create models for: (i) LOF; and (ii) Isolation Forest (IsoF). These are both used because they identify different anomalies.

	- Isolation Forest randomly splits the data, over and over, until each point is isolated. Every point is given a score based inversely on the number of times it took to split the data until that point was isolated. This process is repeated a number of times and an average score is created. Anomalous points tend to need fewer splits and therefore tend to get a higher score. - Local Outlier Factor identifies “local” outliers relative to their neighbourhood. Data points are compared to other data points in this neighbourhood, and given a score based on the density of their neighbourhood, relative to that of their neighbours. A point that is less dense than its neighbours has a higher score. These models capture 'normal' relationships between the variables over the previous 12mths, and units from a selected period are compared to this information. The results were found to be fairly robust to the inclusion of some anomalies in the training data. Other methods would be good to assess, such as classification methods, however more labelled data would be needed. This can be difficult to create where anomalies are few. <u>Variable creation and selection:</u> Regarding the variables used in the LOF model, a number of time-series variables were created in particular to incorporate information about the expected movement for that unit (with respect to itself, or 'like units'). The variable definitions and selection were initially simple - to see whether simple-and-fast to create / understand / maintain models were able to provide a good and robust result. It was found that a relatively small number of appropriately defined variables captured much of the important parameter space needed for fast, low dimension, efficient and effective outcomes. The LOF identifies units with anomalous combinations of these variables. The variables were normalised to incorporate a shift and spread to fix the 5th and 95th percentile (so they were roughly on the same scale, but the tails were allowed to remain long). <u>Hyperparameter selection:</u> The hyperparameters were chosen to be smaller (for faster compute) that still provides good, stable performance, particularly for anomalies. The key hyperparameters were: - LOF: number of nearest neighbours; and - IsoF: number of trees and number of samples to build the tree.

	Performance: The team explored a number of approaches to help determine whether anomalies identified were of interest, and also whether the method was missing key anomalies: - Explored use of visualisations (e.g., time series, scatterplots). - Compared with some key outliers (what would a human consider 'wrong' vs 'unusual'). - Feedback from the business area / data experts. The business areas were very busy, thus it was harder to get their time / input. We also needed to spend time bringing them on a journey. - As the research team and business area learned more about the data, were able to start building a set of known anomalies, which was also useful for assessing the performance of the models. The feasibility assessment was undertaken using samples of data due to memory / processing limits in the environment; and for the visualisations. Random samples were used initially for feasibility assessment. Later work instead used group-specific data/models as specified by the statistical team. This also enabled us to parallelise the preprocessing / training code. We are still learning about the most efficient way to code for dashboards. Learnings: - Categorical variables are problematic for LOF/IsoF, so continuous variables were created to capture the concept, for example by comparing a unit with units in the same category. - Variable normalisation was required, however only basic normalisation was necessary to have good and robust results. - There are some situations where data may have unusually high densities, which can impact the LOF score of nearby units. The current arrangement no longer has this issue, but at the time the issue was dealt with by dropping these units because deemed to be 'boring' (i.e., the same value every period) and so were dropped from the analysis so that they did not impact other units. - Some additional pre-processing was needed. For example, capping very large/tiny values of some ratio variables. - Testing was also undertaken on the various variable definitions, the number of variables and to ensure appropriate targeting (e.g., not identifying large units just because large, or small units just because they tended to be more volatile). Engagement with statistical production area and IT area The team worked with the business area and IT area to build a prototype to demonstrate how the method worked. Multi-level engagement has been important throughout the process, including with the business owners, the IT team, the

	corporate infrastructure funding/management team for this build work, other corporate areas (including data custodians, methods owner). A number of models were assessed (sets of variables) and the key hyperparameter was selected (e.g., nearest neighbours). The team selected a small set of useful models, and these were provided to the business area for assessment and feedback. As expected, the proof of concept showed that LOF/IsoF performed fairly well in identifying anomalies. We were able to get some feedback from the business area throughout the process, which was crucial to ensure the model was useful for their needs. - They are a busy team, thus we needed to be mindful of their availability (e.g., production cycle). - It was important to spend some time over multiple sessions helping the business area become comfortable with the concepts and ideas. We ran some presentations and demos, and also provided them with information and visualisations and allowed them the space to consume and dwell on. - All teams have some turnover so from time-to-time we needed to introduce new staff to the concepts and ideas. - Most of the business area did not have much experience with the environment; were also learning about the tools/environment. - Some of the feedback from the business area related to functionality that needed IT resources to build.
Develop proof of concept	A proof of concept was developed for evaluation however needed additional IT components / processes built and enabled. Some initial IT support - e.g., to build some of the key data analysis / visualisation tools to enable app hosting - was funded and managed by methodology (with goodwill from busy IT area). This enabled the team to build and host an Anomaly detection dashboard for evaluation by the statistical area. This anomaly detection dashboard app / system is currently being built / evaluated.
Approach / method used	A combination of Local Outlier Factor and Isolation Forest was used to identify anomalies. The idea is to send a targeted list of the most significant and most anomalous units to the human decision-makers, along with some contextual information. The data is large and is provided regularly for publication, so fast pre-processing is important. Pre-processing was kept to a minimum and parallelised over subgroups (that are relevant to the outputs and how the decision-makers operate). For every period, the pre-processing extracts the data, creates the variables and the training data/models. The anomaly scores were scaled for interpretability.

	A prototype dashboard was built to enable the human decision-maker to view the short-list of significant+anomalous units (the user can vary the anomaly score and significance score cut-offs; the dashboard compares to the pre-processed models); and view contextual information such as time series plots to help with decision-making. For more detail, please refer to the 'Assess Preliminary Feasibility' section above.
Prepare a Comprehensive Business Case	Future directions will depend on funding.
Deploy the model	Not yet at this stage.
Results	Not yet deployed in production.
Latest status and next steps	Future stages of productionisation depends on funding - so an interim dashboard tool has been built for evaluation by the validation team (currently being assessed - initial feedback is positive), with some IT components/systems currently being built to enable business areas to host their own dashboards. Upcoming work to: - Incorporate feedback from evaluation. - Work with statistical production team on maintenance of system (when and how), including work on explainability. - Investigate application to other statistical products. - Assess feasibility of automated treatment of less-significant anomalies. - Any future productionisation stages will need to consider testing, tech debt, etc.
Lessons learned & recommendation	Learnings included: - Machine learning can provide benefits for anomaly detection, including finding unexpected anomalies, better targeting lists of anomalies sent to validation teams, providing more contextual information for the validation team, managing large datasets, leveraging multi-variate analysis. - It was found that a relatively small number of appropriately defined variables captured much of the important parameter space needed for fast, low dimension, efficient and effective outcomes. A set of group-specific models was developed that suited the statistical production team and for efficient processing and use within the dashboard tool.

	- Initially aimed for low-risk / easy to explain machine learning solution … however once production areas were comfortable with it, they very quickly wanted more-advanced approaches. - New compute environments offer opportunities for big data and new approaches, however a lot of effort, for example: - Some additional functionality / tools become available, but other usual functionality is not/yet available. Particularly the case for research environments. - For business areas who work in these new compute environments there is a very large amount of additional knowledge that is needed. - There is a very large amount of IT effort to build environments suitable for business areas to have greater control over their own statistical products. e.g., building environments / systems for business areas to create their own apps takes work, and business areas need to build / maintain a different set of skills. - It was very helpful to have some staff who could 'bridge the language gap' between IT / not-IT areas. - Close connections with the business areas were crucial. Found it was important to provide useful concrete outcomes along the way. For example, the team identified some key anomalies and proposed some interim rules to help the validation team identify anomalies in the short term. Ongoing engagement with IT teams, Methodology support areas and other corporate areas were also very important.
Reference	No publications available yet.

Use case 2	**StatCan - Unit Value (UV) Error Detection and Correction: A Machine Learning Approach**
Project overview	In short, the goal of the work is to improve the quality of import data, specifically the Quantity and Unit Value (UV) fields, received from administrative sources. These data are used to produce indicators on international trade (import statistics) as part of the system of macroeconomic accounts. The issue is that the UV (or Quantity) is often misreported on Customs Declaration forms. The UV is a derived variable from the reported Quantity and Value. The Value is carefully checked by Customs agents but the Quantity field (and thus the Unit Value field) less so. This results in: ● A great deal of User Inquiries ● Significant time investment in the review process by data processing/production analysts An "edit and imputation" approach to detect and correct/impute erroneous Quantity/UV fields exists but was determined to be underperforming and inadequate. The existing approach largely focuses on "clipping" extreme values and imputes them with random donors. Given the large size of the micro dataset (import declarations) there is little room for manual validation and edit. The business need of the project was to develop a new error detection and imputation approach (as not all extreme values are errors and not all errors are extreme values). A machine learning model approach was chosen for exploration as such methods had not been tested in the past and were known to show promise for processing with large data sets. Work mostly started in 2019 (exploration started in 2017). Work now (fall 2022) is at the final steps of implementation.
Organisational readiness	Low to moderate: ● Initially, the tools (e.g., ready access to python/R packages) and compute infrastructure were initially lacking. • IT Service Providers within the organisation were very hesitant to provide broad access to open-source tools out of a fear that more cyber security breaches could occur (in addition to not yet having a clear process to maintain and support such tools). • Over the 2-3 years of the work, this changed as higher performance machines were purchased, cloud access increased, and clear product owners for the needed open-source tools were identified.

	• Expertise in ML was still at early stages when the project started in 2017 (e.g., a handful of employees had any experience and were found 2-3 teams). • In 2022, there were potentially 50+ that have experience in the ML methods being applied with a large concentration in the Methodology Team but also with a few small but important pockets with the Subject Matter teams. • Organisationally, there was supported to try new methods.
Understand business needs (Who needs what)	The business need was clear: The current E&I approach was underperforming. Desire for a new approach was high. Although a new method for E&I was being explored, that was not the goal. The goal was always to develop a better E&I method.
Assess Preliminary Feasibility	The choice to try an ML approach can in part be attributed to the large size of the micro dataset, large number of different products and large variability. Trying to come up with a successful rules-based approach was unlikely. Given a "clipping" approach was already in use, trying something new was warranted and ML approaches were showing promise on large datasets. Initial feasibility of an ML work was low as the quality of existing labelled/training data was low to test a machine learning (ML) approach. Initial ML model performance (with an XGBoost-based model tested on non-representative samples) was promising but not particularly good. Much work went into improving the quality of the existing labelled data. On an ongoing basis, production staff labelled random samples of the new months for use as validation / testing data. "Business rules" were used to correct errors in historical training data. Once done, model performance was tested and found to be good, significantly outperforming the existing E&I approach.
Develop proof of concept	The goal was to develop a new E&I method. Initial exploration showed an XGBoost-based model approach seemed promising. The key thing to mention is that the development of the Proof of Concept (test an ML approach) came out naturally from the business need for a better performing E&I method. It was not the reverse (e.g., find a use case to test an ML approach). An important obstacle at early stages was the low quality of the initial training data. If the signal is bad, no amount of modelling (no matter how advanced) is likely to make any sense of it. Improving the quality and of labelled data was key. This has shown true in subsequent application of machine learning as well.

Prepare a Comprehensive Business Case	The business needed an improved system for E&I of unit value and wanted this project to go ahead. A key point is that the need was not for a modern method. The need was better results. Such business cases are standard and don't really require anything special or new.
Deploy the model	The model deployment was done through a production-ready Windows R server accessible by the rest of the International Trade programmes and system. The deployment of the model went fairly smoothly. Much discussion occurred with IT partners to come up with the best way to deploy. Having and obtaining the needed R-servers did take some time but this occurred in parallel over the course of the project as part of a broader initiative to have R/Python servers that could be used for official production.
Results	The model has not yet been deployed in production. The model results have been tested with various metrics on held-out test data from new months not seen by the model. The results were checked in a quality assurance (QA) dashboard developed by an independent team. Metrics checked include: ● Mean absolute error (MAE). ● Mean squared error (MSE). ● Various classification-based metrics (like fraction of points incorrectly edited when should have been kept as-is, fraction of points kept as-is correctly, etc.). The model results show significant improvements over the existing system (which was chosen as the benchmark) in a wide variety of product categories. The ML approach performed as well or better than the current E&I approach (UV-Clipping) in all product categories.
Latest status and next steps	The model and project are in its final stages of approval. ● The methodology has been approved and reviewed for appropriateness and potential ethical consideration. ● A code review was completed to help ensure the robustness of the model code (e.g., it is easily maintainable and well documented). ● A formal Service Level Agreement that defines production roles and responsibilities across Production Staff, IT Staff, Data Scientists, and Statisticians has been developed and signed. ● The work to deploy the model and integrate it into the current statistical processing systems/flow is complete. ● The key remaining task is to brief senior management and train production staff on how to use the QA tools that have been developed to monitor the performance of the models.

	After this, the assumption is that the full transition to production (switching over from the existing E&I system to the new system) will be done.
Lessons learned & recommendation	● Execution: It is important to first have results from the proof of concept before doing any transition to production work. Some work on making the code "production-ready" was done before we had quality results and it turned out to be a waste of time given changes made later (which were verified on new random samples from newer months). First get the proof-of-concept results working well before working on any transition to production. ● Execution: In addition, do not make the code too general and "heavy" before getting quality results. Work was done on having the code work for additional variables and support for other algorithms, which turned out to be unused and was deleted later. The code has a specific business objective and should focus on that, it is not a generalised system for doing lots of different things. ● IT Infrastructure: Unless a deployment plan is known from the start, begin working early on with IT infrastructure providers to understand what a likely deployment architecture might look like. If it isn't available acquiring the needed infrastructure for production deployment can take some time. ● Organisation knowledge of the proposed method: Given the methods are unlikely to be known by many, don't get caught up in traditional responsibilities of services provision and maintenance. Pool the expertise that does exist and work as a multi-disciplinary matrix team.
Reference	No public resources are available at this time.

<table>
<tr><td>Use Case 3</td><td>INE (PT) – Anomalies detection and imputation on administrative data</td></tr>
<tr><td>Project overview</td><td>Statistics Portugal (INE) started to analyse several approaches for anomaly detection and imputation of data from enterprise invoices (aggregated by enterprise and buyer) provided monthly by the Portuguese Tax Authority.

INE receives this administrative data on a month m with data referenced to m-1, on average around 85 million records per month, covering around 1 million different sellers. The data structure is as follows:

<table>
<tr><th>Year</th><th>Month</th><th>Seller</th><th>Buyer</th><th>Value (€)</th></tr>
<tr><td>2022</td><td>8</td><td>seller1</td><td>buyer1</td><td>204,35</td></tr>
<tr><td>2022</td><td>8</td><td>seller1</td><td>buyer3</td><td>1154,12</td></tr>
<tr><td>2022</td><td>8</td><td>seller1</td><td>buyer4</td><td>115,33</td></tr>
</table>
There are some issues with the data, as it may have insufficient coverage depending on the day of the month the data is extracted by the tax authority.</td></tr>
</table>

Organisational readiness	The analysis and treatment of administrative data is following a new INE strategy of centralised data process and treatment: one dataset serves different users. To achieve this purpose, since 2019, some adjustments have been made in the internal organisation for strengthening the capacity for data management and analysis in two departments: Methodology and Information System Department and Management and Data Collection Department. Meanwhile, INE management has prepared and encouraged training courses in data science tools, both to empower employees with new knowledge and to bring machine learning methods both to empower employees with new tools and to bring machine learning methods into their daily work activities. In the middle of 2020, a new unit was created (Administrative Data Unit, under the umbrella of Data Collection Department), responsible for: ● Evaluation and testing the use of new data sources, with a view to improving the quality and consistency of statistical production; ● Evaluation of the possibility of replacement of the information collected by surveys or censuses; ● Definition of new validation models, consistency, and coherence analysis ● Integration of data from various sources
Understand business needs (Who needs what)	For these administrative data to become statistical data, it must be treated and validated, to ensure quality, reliability, consistency, and completeness of the data. In this data cleansing process, we also perform a more in-depth and specific analysis of content handling anomalous or lack of information. Although this dataset serves many different users, it has a user group that is very interested in the success of this process: the short-term statistics team. In this case, we have brought the colleagues of this team to frequent meetings where we inform them of the progress and setbacks in the identification of anomalies and their treatment, presenting them with possible solutions and results and

	showing openness to their contributions and possible proposals. The users need to have the data available to work within 2 working days. The continuous improvements in the treatment process have allowed the data to be delivered around 30 hours after its transmission.
Assess Preliminary Feasibility	No preliminary feasibility study has been developed, but evolutionary and phased work has been done, involving users, to make the results more robust and accepted by all. During this process some analysis tools and approaches were used such as: ● Data exploration using time series visualisation; ● Comparison of the results obtained with survey and extrapolated data; ● Comparison of the historical data with annual reported data; ● Knowledge and feedback from key users about the potential anomalies identified (some of which could have an explanation).
Develop proof of concept	Identification of missing values and its imputation applied to the monthly taxable amount of a small but sufficiently relevant set of units, capable of ensuring a remarkable quality improvement in the data processed.
Prepare a Comprehensive Business Case	A solution is needed to solve the problems encountered when the e-invoice data received does not have sufficient coverage (have many missing values). The model must discriminate between total missing values and partial missing values (abnormally low values and records).
Deploy the model	The process, deployed in R language, is based on the following R-packages: ● {tidyverse} for data manipulation, ● {targets} for defining a workflow for functional programming, ● {isotree} - Fast and multithreaded implementation of Isolation Forest (a.k.a. iForest) for anomaly detection ● {imputeTS} for imputing missings in univariate time series,

- {ROracle} to create an interface between R and Oracle database,
- {tsibble} for time data manipulation,
- {fable} and {fabletools} which provide forecasting models for time series,
- {RJDemetra} interface for seasonal adjustment software officially recommended for members of the ESS,
- {Metrics} for the implementation of validation metrics used in supervised machine learning methods.

In order to evaluate the best nowcasting method, the following models were applied to each of the seller series:

- ETS - Exponential smoothing state space model, the best model is chosen automatically;
- ARIMA - a variation of the Hyndman-Khandakar algorithm is applied to obtain the best ARIMA model;
- NNETAR - Neural network autoregression, fits a NNAR(p,P,k)m model with a hidden layer.
- Prophet - fully automated facebook forecasting procedure;
- X13 - X13-ARIMA method for estimating seasonal adjustment of time series;
- TRAMOS - TRAMO-SEATS method for estimating the seasonal adjustment of time series;

For validation and selection of the models, data from January 2016 to December 2021 was used as training and for testing, data from January to May 2022. The results obtained from each of the models for the test data were compared with the "real" values through validation metrics like RMSE, MAPE and MASE. Historical time series, for each one of the relevant sellers, were corrected from isolated missing values with Kalman-Smoothing method or by applying the chain variations of the respective NACE activities.

The procedure is now running monthly for the identification of missing values. For those time series with missing values, it is selected as the best model for nowcasting. The imputed anomalous values are integrated in the database to be made available to users, with the proper identification of the imputation made.

Results	The feedback from the users (in particular the short-term statistics team), on this dataset treatment has been very positive. The values obtained after the treatment of anomalies have been compared with the values obtained through survey and are much closer than the original values received from tax authority.
Latest status and next steps	We consider our approach to be conservative but robust as it is based on analysis and imputation of large enterprises which, while they may have diversified behavioural patterns, offer some guarantee of stability. So, because our focus was on large companies, there are still some issues to resolve among all the other vendors. Due to the high number of companies involved, we think we will be obliged to use different approaches, according to the different characteristics of enterprises. We are also awaiting final versions of the data from the tax authority which will allow a more accurate assessment of the results obtained. Looking for and testing new methods for nowcasting, for instance, an ensemble method.
Lessons learned & recommendation	The involvement of colleagues either from the methodology department and from the accounts department (responsible for the STS), has been crucial for a sustained and credible advancement of any process related with data quality improvement.
Reference	No public resources available currently.

Use case 4	**SFSO – Imputation using missForest**
Project overview	In 2018 an external mandate showed unsatisfactory results for the imputation of fortune variables in the Survey on Income and Living Conditions (SILC) using the IVEware software for the fortune module. The main problem was that distributional accuracy could not be achieved. However, the distribution of the variables is of high interest in this context because the results are used in poverty indexes. Slightly better results could be achieved with knn. These findings encouraged SFSO's Statistical methods unit to investigate the quality of the missForest algorithm in a simulation framework and extend it to material and social deprivation variables. A further extension of the simulation tests to income variables has also been decided.
Organisational readiness	The organisation was ready for the change with respect to - the infrastructure, - openness and - the needed skills.
Understand business needs (Who needs what)	The overall business need was clear: The current E&I approach was underperforming. Desire for a new approach was high. The goal was always to develop a better E&I strategy. Therefore, the aim was to quickly gain an insight on the feasibility and the quality of using missForest. Hence, it was decided to test missForest for the smallest set of variables (fortune) first and extend the tests to material and social deprivation afterwards because of the relatively high amount of missing item non-response and few relevant auxiliary variables at hand for these variables. Only after that, the income variables, which concern by far the biggest number of variables, with the highest item non-response rate, were considered in the testing. However, these last imputations could have an effect on the first two modules and if the imputation of income variables will be successful it is advised to re-run those for the material and social deprivation variables and those for the fortune variables. Based on our understanding, the filtering questions unfortunately prevent the imputation of all variables at the same time. The choice of this strategy was also influenced by the available resources.
Approach/method used	The approach of evaluating the performance or the algorithms consisted in a simulation framework where missing values were generated based on the missingness mechanism observed in the survey data.

	Knn was used at a preliminary stage. Finally, missForest was used due to better performance than knn.
Assess Preliminary Feasibility	Based on the fact that the data set is not very large, about 7'300 households and due to filtering, there were only between 2'200 and 5'600 households concerned by the fortune variables with an item non-response rate between 10% and 15% it was not sure that a ML algorithm would be appropriated. The same problem occurred for the 13900 persons in the net sample for material and social deprivation variables and an item non-response rate of about 18%. However, the simulations showed encouraging results in both cases. We had also to take into account a questionnaire redesign for the fortune variables (splitting of variables and added range responses) in the simulation of the fortune variables as those real data were not available at that time. Furthermore, due to a lot of true zeros for some variables in the fortune module, it was necessary to add an imputation based on a logistic regression to get rid of these zeros. Otherwise, an important part of the imputed values was outside the range values observed and the distributions of those variables were distorted.
Develop proof of concept	The proof of concept consisted in the simulation framework.
Prepare a Comprehensive Business Case	The setup of the simulation tests accounting for a questionnaire redesign (splitting of variables and added range responses) showed to be a realistic and a comprehensive business case. The random generation of missingness patterns based on the observed ones needed a lot of resources.
Deploy the model	The model deployment consists in integrating the R-code into a SAS production pipeline.
Results	The models have not yet been deployed in production. Validation on the generated missing values sub-sample has been done by observing. • Mean absolute error (MAE, called total error in the documentation above). • Main error: same as MAE but limiting the error to a change between material deprivation. • Confusion matrix. • Decile boxplots of the error distribution. • Imputation impact (based on imputing the missing values on the real net sample).

	The results show significant improvements over the existing system (which was chosen as the benchmark). The impact on the distribution of the variables of interest and derived indexes showed encouraging results for the fortune module and the material and social deprivation variables.
Latest status and next steps	• The simulation study for the income variables is still going on and has to be finished. • The validation of the results of the imputation of the income variables by domain experts needs to be done. This step also includes the assessment of the impact on already published results. • Based on the results of the imputation of the income variables, the fortune variables and the variables on material and social deprivation should be re-imputed. • Based on the assessment of the impact on the results of the income variables it has to be decided how to handle time series and how to organise the communication with the general public and with stakeholders. • A formal decision by the general management based on the above-mentioned items might be necessary to implement the missForest imputation algorithm into production. • It is planned that these imputation tests will be documented in a methodological report.
Lessons learned & recommendation	• Execution: A thorough validation based on a simulation framework is very time consuming. This has to be clear from the beginning. • Execution: The transition from simulation tests from one variable set to another is not straightforward and is also time consuming. • IT Infrastructure: no issue so far. • Organisation knowledge of the proposed method: no issue so far.
Reference	For the simulation tests of the material and social deprivation variables, see https://unece.org/sites/default/files/2022-10/SDE2022_S4_Switzerland_Bianchi_AD.pdf. Otherwise, there is no public documentation available at the moment.

Use case 5	**Statistics Sweden - Imputation of Occupation in the Occupational Register**
Project overview	The Swedish statistics on Occupation come from the Occupational Register, which contains information on the occupation of individuals. The occupational information is intermittently collected from businesses, and is therefore subject to missing values, especially for younger and older individuals. Imputation of occupational information can reduce the proportion of missing values. The current model for imputation of Occupation is becoming obsolete and a new model needs to be developed. In addition, the population for occupational statistics is to be expanded, which may increase the number of missing values. To address this, Statistics Sweden has developed a machine learning model for imputation of Occupation. The model uses register variables on the individual level and the employer level to predict Occupation. The development of a machine learning model for imputation follows the strategic and operative goals of Statistics Sweden, which emphasises the use of machine learning for automated methods such as imputation.
Organisational readiness	The organisational readiness of Statistics Sweden is varying. The expertise on statistical methodology and data science to develop machine learning models is good. The machine learning IT infrastructure is less developed. Statistics Sweden has developed a process on development and implementation of ML methods. The process is accessible in the statistical production system of Statistics Sweden. Further development of the process includes additional process steps on assessing business needs, quality requirements, and prerequisites, and on the monitoring of ML models. The process may be used to support the development of machine learning models.
Understand business needs (Who needs what)	The project was initiated by subject matter experts for the Occupational Register. The aim of the project is to replace the outdated imputation model with a new model. Imputation is needed to address the issue with missing values in the Occupational Register. If the imputed values have the same quality as the other observations in the register, the quality of the statistics will increase.
Assess Preliminary Feasibility	The model utilises several register variables to predict Occupation. It is likely that traditional imputation methods would be less successful in realising the potential of the auxiliary information to predict Occupation; hence, it was

	decided at the initiation of the project to use a machine learning approach. This is also in line with the strategy of Statistics Sweden. We considered only tree-based methods, i.e., random forest and gradient boosting, because such methods have shown good performance on similar problems previously.
Develop proof of concept	The development of a proof of concept was integrated in the development and was made during the early stages of the development. Because the predictive performance of the model was lower than stakeholders expected, it was decided that we should aim to impute Occupation to facilitate the production of statistics instead of aiming for individual level accuracy.
Approach/method used	The model was trained on data from the 2019 Occupation Register on the gainfully employed population 16-74 years old. Features were extracted from the variables in the register. The random forest model was used because it showed similar predictive performance as the gradient boosting model and needed less resources for training. Evaluation of the model was done with respect to individual predictive performance, class level predictive performance, and the effects on the statistics. The individual predictive performance was evaluated using accuracy, precision, recall, and F1. The class level predictive performance was evaluated by simulating the missing data mechanism in validation data and replacing simulated missing values with imputed values, which facilitated the joint evaluation of the missing data mechanism and the quality of the imputed values. The effects on the statistics were also evaluated using simulated missing values and by imputing values on previously missing data and considering the effect on the distribution of Occupation.
Prepare a Comprehensive Business Case	The business case was successful because the task was clearly formulated from the outset. However, modifications had to be made with respect to the expected outcome and performance of the model.
Deploy the model	The model is yet to be deployed in production.
Results	The model is yet to be deployed in production.
Latest status and next steps	The project is currently in the deployment phase.
Lessons learned & recommendation	The project has highlighted the need for a process to facilitate the assessment of business needs, quality requirements, and prerequisites. If such a process had been in place, it would have been clear from the outset how to proceed with respect to the measured performance of the model. In addition, the project would have benefitted from further clarification of the expected

	use of the imputed values.
Reference	ML2023_S1_Sweden_Malmros_A.pdf | UNECE

Use case 6	**Bank for International Settlements - Time Series Outlier Detection using Metadata and Data Machine Learning in Statistical Production** **Organisational aspects of implementing ML based data editing in statistical production**
Project overview	The BIS Data Bank is a data warehouse hosting more than sixty thousand macroeconomic and financial time series. Data quality checks currently in place in the BIS Data Bank identify outliers relying on traditional statistical methods (e.g., standard deviation band). These methods are typically based on predefined thresholds which may not be suited for time series with linear breaks, such as financial time series. Furthermore, it does not allow for contextual outlier detection (e.g., using cross-country data for the same indicator which is largely available in the BIS DataBank). We propose a new method relying on machine learning that performs outlier detection taking into account also related time series. Our method has two main steps. First, time series are clustered based on their metadata and data. Second, contextual outlier detection is performed for each cluster. Our proposal aims to improve the current statistical production pipeline for the BIS Data Bank.
Organisational readiness	As the new method is not deployed in a production pipeline yet, it did not require specific organisational arrangement. However, synergies between IT and business teams are key to facilitate the deployment of innovative solutions, mostly of which are already available at the BIS (e.g., Python workbench, connectors to access the data, Azure DevOps).
Understand business needs (Who needs what)	The BIS Data Bank is undergoing a migration process. A reshuffle of the current in-house FAME-based software is ongoing towards a Python-based solution to perform most of the tasks covered by the Generic Statistical Business Process Model (GSBPM). The goal is to improve the overall efficiency of the existing statistical pipelines (e.g., less manual intervention, better DQM). The new ML-based outlier check could be leveraged in this context
Assess Preliminary Feasibility	The early stages of the project include an in-depth comparison of the new method against the current one, with a focus on accuracy/data quality. Other key aspects are optimisation of manual intervention and domain-specific knowledge (e.g., for the choice of ML algorithms), generalisation of the model (e.g.

	to micro/unstructured data), code transparency, black-box and lock-in issues. For the full development of the PoC other considerations will be required: performance, ML pipeline setup
Develop proof of concept	After the initial assessment feasibility, we aim at delivering the Proof of Concept on a limited but composite sample of the BIS DataBank and benchmark it against the current checks. This stage will require more rigorous tuning of the algorithm and check its performance.
Approach/method used	To prototype our method, we plan to test the accuracy of the model against multiple data types (indexes/prices, stock/positions, flows/transactions), parameters and pre-processing techniques (e.g., scaling cannot be applied across all data types). We will also tune the frequency of the checks against the update frequency of the underlying data and test the performance.
Prepare a Comprehensive Business Case	At this stage, the main driver of the project is to provide a better solution to increase productivity and reduce manual intervention on DQ checks.
Deploy the model	Not applicable
Results	Not applicable
Latest status and next steps	The method is not in production yet. The next stage is to further enhance the outlier detection algorithm and develop a Proof of Concept.
Lessons learned & recommendation	Not applicable
Reference	UNECE Machine Learning for Official Statistics Workshop 2023 | UNECE.

Use Case 7	**Statistics Spain (INE):** **Early Estimates of the Industrial Turnover Index using Statistical Learning Algorithms**
Project overview	The final aim of this project is to obtain early estimates of the Industrial Turnover Index (ITI) even before finishing the data collection and data editing processes, thus improving the timeliness but keeping the accuracy of the early estimation under control. Currently, the dissemination of the index is carried out around 51 days after finishing the monthly reference period. However, the response rate is around 75% 21 days after finishing the monthly reference period. So, it was considered to explore new methods to provide more timely information. These new methods amount to performing fine-tuned mass imputation in the microdata set for those sampling units not yet collected. This way, the index is obtained combining the units already collected and edited together with the imputed values. The estimation error is also computed. Collaboration with the subject matter experts is essential to include highly relevant information into the estimation process and how to deal with some issues that are raised during the project. The pilot prototype was developed in 12 months, and it is already finished.
Organisational readiness	Statistics Spain is open to ideas regarding modernisation and innovation. The organisation provides the possibility to set up collaborations among different units such as (IT, methodology, and domain experts). There also exist several internal working groups about specific issues such as seasonal adjustment, National Accounts and short-term business statistics, temporal disaggregation, etc. There is also an important amount of specialised knowledge personnel with good expertise in their specific areas. However, regular production of official statistics according to the National Statistical Plan and the European Statistics Programme constitutes the top priority, thus activities are strongly oriented towards this goal so that it is challenging to

	modify or introduce novelties in the statistical production processes. This also entails challenges and non-negligible efforts to implement and maintain new statistical products. The main challenges to deploy new proposals can be shortly summarised in (i) the lack of some professional roles or skills regarding Machine Learning techniques at an institutional scale and b) the lack of computational resources and structures appropriate for the execution of new computationally demanding methods at an institutional scale. These challenges are increasingly tackled with measures such as the organisation of internal courses about programming languages for modern data analysis techniques and the deployment of centralised computational facilities with these languages.
Understand business needs (Who needs what)	There is a huge need for improving timeliness in the production of official statistics. Short-term economic statistics are especially relevant to obtain fast economic indicators. Then, having early estimates of the industrial turnover index and similar short-term business statistics is relevant both for internal users such as National Accounts Departments and for external users and stakeholders as well. Furthermore, the need for timely information has become extremely obvious in recent times of uncertainty under a global pandemic.
Assess Preliminary Feasibility	Some assessments were made at the beginning of the project to evaluate the viability of this product in terms of quality, especially timeliness. The idea of performing imputation using machine learning techniques was clear from the first moment due to the versatility and predictive power of these methods. However, some preliminary analyses were made to choose the best model for the specific problem at hand, namely, both the target variable and most of the regressors are continuous. After trying different models (with some preliminary testing), a gradient boosting algorithm was chosen. In order to develop the proof of concept, the available resources (both IT and human resources) were tight. Sound methodology and good-enough accuracy was primed over fine-tuned models to gain in time and to save in computational demands. There was not a detailed evaluation

	in advance of all the required resources and their availability to deploy the pilot study in production because the priority was to assess the viability of the underlying ideas and the general approach. The developers of the prototype worked with PCs implementing the source code in R language. Expertise in ML techniques has been gradually improved thanks to the participation in international projects.
Develop proof of concept	The development of the proof of concept was carried out with real survey data of the Spanish Turnover Index from Oct17 to Dec 2021. For each successive month, the statistical model was trained with data from the past time series and applied in turn to the reference time period, of course emulating real-life production conditions. Accuracy was assessed compared with real validated data from the survey. Notice that predicted values can always be compared to real validated values after the whole survey compilation and execution is over. The model and estimates are continuously updated when data is made available to domain experts from the data collection and data editing stages. This process was executed in a batch for 60 consecutive months. At this point, the subject matter expert knowledge was recognised as fundamental for the information representation step. Feature engineering incorporated most of this knowledge. After encoding 287 regressors were built based on 10 variables using both the reference period values and historical values. Interestingly enough, to compute the estimation error and to cope with the different statistical behaviour of sampling units (business populations are highly skewed), the exchangeability hypothesis was dropped, introducing some changes in the standard computation of prediction errors with these techniques. The pilot implementation was refactored to allow for an iterative incremental computation and updating of the time series, thus bringing the pilot closer to production. Iterations can run parallel to data collection conditions (daily, weekly…). The increase in complexity is justified because of

	the versatility and adaptability to real-life production conditions.
Prepare a Comprehensive Business Case	In this project the proof of concept was done with a comprehensive business case. The domain expert team was involved and collaborative all along. They provided all the needed data, knowledge, and subject-matter support. They were involved in the project, and they were aware of the implications concerning the great improvement in quality.
Deploy the model	The project is potentially ready for deployment in production using the development code, which is not optimal and provides room for noticeable refactoring (memory usage, I/O optimisation, etc.). Thus, it is preferable and advisable to revise the implementation to be adapted to a MLOps platform connected to the data collection process. Model optimisation through hyperparameters fine-tuning, regularisation and other model selection techniques is also advised.
Results	The model has not yet been deployed in production. The results of the proof of concept are showed in a Shiny dashboard: https://sandra-ba.shinyapps.io/Advanced_ITI_indices_v1/ The development code is available and shared in Github: https://github.com/david-salgado/AdvITI
Latest status and next steps	Nowadays, the project is in pause waiting for the resources to take the leap to production. There is a full development prototype implemented which could be used to publish the pilot as an experimental statistics. Nonetheless, results so far have triggered complementary methodological considerations regarding response burden reduction, non-response treatment, and imputation beyond the sample (in the population frame). In collaboration with two Spanish Universities, the next steps to be carried out in the following years will be to revise the machine learning methods as well as the hyperparameters to try to improve the current results. New regressors will be defined and new data sources will be included in the project.

	Finally new uses of the mass imputation of the microdata set will be analysed.
Lessons learned & recommendation	Methodology of statistical production: The use of statistical learning methods can clearly streamline business functions to improve quality dimensions by reorganising the production process. IT infrastructure and capacity: The availability of a computational platform and the human resources with the required computing skills are needed both for development and for production. Organisational knowledge of the proposed method: It is important that the product is spread not only to the external users but also internally. In this case, we have done a working paper to share the details. Maintenance of the method once in production: We highly recommend that the unit in charge of the support in production, is planned ahead and involved from the first steps of the project. Acceptance of the method by business areas: Since the collaboration with subject matter experts has been established from the beginning of the project, the acceptance and support are reached, and they are aware of the importance and the need of the new product. Their contribution has been essential in the development to overcome difficulties.
Reference	A working paper with a full description and some results is published in the INE webpage: https://www.ine.es/ss/Satellite?c=INEDocTrabajo_C&p=1254735116586&pagename=ProductosYServicios%2FPYSLayout&cid=1259953795823&L=1 Second Position of the 2022 IAOS Prize for Young Statisticians: https://www.iaos-isi.org/index.php/statistics-prize