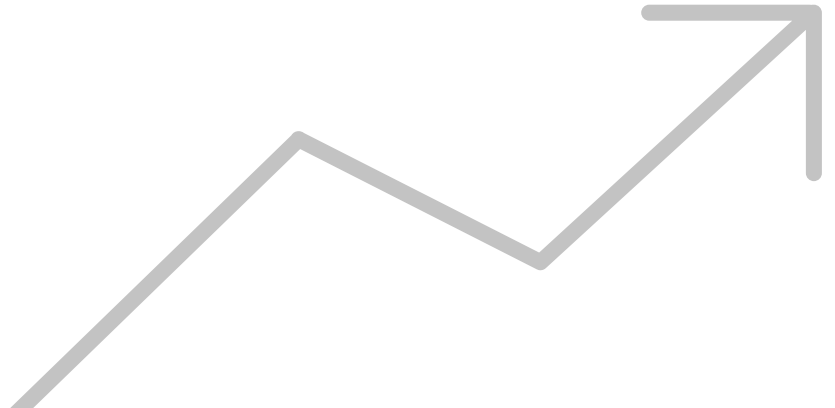
DESTATIS
Statistisches Bundesamt

# Organisational Aspects of Implementing ML Based Data Editing in Statistical Production

## UNECE Expert Meeting on Statistical Data Editing 2024, Vienna

Steffen Moritz (German Federal Statistical Office)



# Background Data Editing Task Team

- This talk presents the results of the 2023 ADSaMM Data Editing Task Team
- The UNECE Applying Data Science and Modern Methods Group (ADSaMM) assembles yearly task teams to identify opportunities to modernize NSO's business processes
- Task Team Data Editing: "Accelerating the Implementation of ML-based Solution in Data Editing"

ADSaMM Data Editing Task Team 2023:

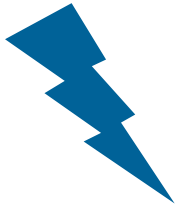
- Claire Clarke (chair) and Jenny Pocknee Australian Bureau of Statistics
- Wesley Yung, Jean Le Moullec and Stan Hatko Statistics Canada
- Riita Piela – Statistics Finland
- Steffen Moritz – German Federal Statistical Office
- João Poças –Statistics Portugal
- Sandra Barragán, David Salgado and Elena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

# Task Team Data Editing: “Accelerating the Implementation of ML-based Solution in Data Editing”

## 2019/2020

*Data editing and imputation promising use cases to apply machine learning*

Report of The Machine Learning project by UNECE High-Level Group for the Modernisation of Official Statistics



## 2023

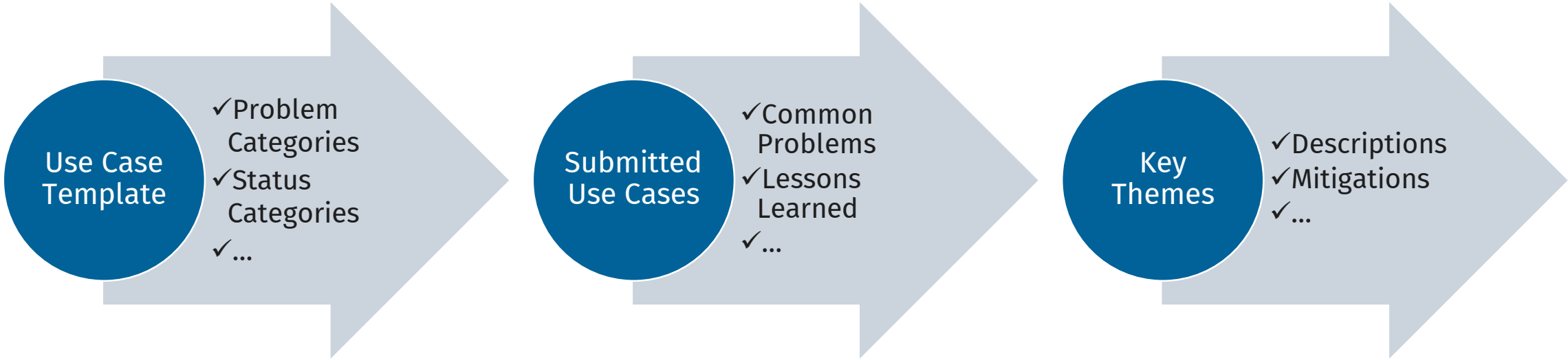
*NSIs have been rather slow to adopt machine learning methods for editing*

Self-reported

# What are organisational blockers preventing the adoption of ML methods and how can we possibly overcome them?

**Organisational Aspects of Implementing ML Based Data Editing in Statistical Production**

# Process taken by the task team



# Use Case Template

- Title of the use case and the name of the organisation
- Project overview
- Organisational readiness
- Understand business needs (Who needs what)
- Assess Preliminary Feasibility
- Develop proof of concept
- Approach/method used
- Prepare a Comprehensive Business Case
- Deploy the model
- Results
- Latest status and next steps
- Lessons learned & recommendation
- Reference
- Contact

# Evaluated Use Cases

- 7 use cases
- Quite different projects
- In different stages of editing process
- Mostly not in production yet

| |
|---|
| Un-supervised ML for anomaly detection in large and frequent admin data (Australian Bureau of Statistics) |
| Unit Value (UV) Error Detection and Correction: A Machine Learning Approach (Statistics Canada) |
| Anomalies detection and imputation on administrative data (Statistics Portugal) |
| Imputation using missForest (Swiss Federal Statistical Office) |
| Imputation of Occupation in the Occupational Register (Statistics Sweden) |
| Time Series Outlier Detection using Metadata and Data Machine Learning in Statistical Production (Bank for International Settlements) |
| Early Estimates of the Industrial Turnover Index using Statistical Learning Algorithms (Statistics Spain) |

# Identified Key Themes

## Key organisational themes influencing ML adaption in data editing

A. The driver of the problem being addressed by a machine learning solution

B. The lack of labelled data or other suitable training data

C. The relationship between business area, methodologists/data science staff and IT specialists

D. The need for input and feedback from subject matter specialists

E. Domain specific knowledge and the black box nature of machine learning methods

F. IT issues and Machine Learning Operations and machine learning platform

# A. Driver of the problem

What is the motivation behind the introduction of AI/ML solutions?

Changing established approaches requires significant motivation

- Often greater trust in human-led processes
- Satisfaction with the quality of the current methods

‘*If it isn’t broken don’t fix it*’ attitude from business sections

# Suggested Mitigations

## Driver of the problem

- Crucial to find the right opportunities to introduce ML for editing
  - Acquisition of new data (e.g. high volume, high frequency)
  - Improvement to current methods (e.g. current editing solution is clearly inadequate)
  - New products or services

- Reframing problems may help present them as suitable for ML solutions

# B. Lack of training/labelled data

Are the data prerequisites for ML solutions in place?

- Lack of (high quality) **labels**
  - Absence of labels
  - Low quality /biased labels
  - Delayed availability of labels
- Lack (quality, amount) of **training data**
  - Not enough data
  - Non-representative data
  - Delayed availability of data

# Suggested Mitigations

## Lack of training/labelled data

- **Use Human Expertise**: Involve human reviewers for label generation and critical edits.
- **Speed Up Label Availability:** Work on accelerating label and data availability to improve model integration.
- **Improve Label Quality**: Focus on enhancing label quality through consensus analysis and uncertainty quantification.
- **Address Sampling Bias:** Ensure training data is representative and mitigate bias through statistical methods.
- **Use Simulation Studies:** Apply overimputation and simulations to evaluate missing data.
- **Leverage Partial Data Deliveries:** Work with partial data and label deliveries while improving data availability
- **…**

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

Who is responsable for what? How to work together?

- Collaboration between **methodology** and **data science**
- Collaboration between **business experts** and **data science**
- Collaboration between **IT** and **data science**

*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.*

# D. Input and feedback from subject matter experts

## How to get the relevant information from subject experts?

- **Limited Time**: Subject experts often need to focus on urgent production issues, limiting innovation involvement
- **Scalable Solutions**: To develop standard solutions data scientists need input from multiple subject areas / experts
- **Steep learning curves:** Subject experts face steep learning curves understanding new methods and their requirements

# Suggested Mitigations

## Input and feedback from subject matter experts

- **Early Involvement**: Involve subject matter experts from the start and recognize their key role.
- **Understand Needs:** Understand their needs early and design solutions accordingly.
- **Training Support**: Provide training to ensure they are comfortable with new methods.
- **Long-term Benefits**: Highlight long-term benefits and time savings to encourage participation.
- **Collaboration Model**: Use a hub-and-spoke model for collaboration.

# E. Requirements for data science expertise and black box issues

Are we having ML experts in-house? Are ML methods accepted in-house?

- **Lack of ML Expertise**: Many organizations struggle with insufficient in-house knowledge of data science and machine learning techniques.
- **Need for Continuous Training**: Staff training and recruitment are crucial to keep up with the rapid advances in machine learning.
- **Transparency and Black-Box Risks**: ML methods can be opaque, posing challenges in explaining unexpected outcomes, which increases operational risks.

# Suggested Mitigations

## Requirements for data science expertise and black box issues

- **Team Collaboration**: Ensure close cooperation between data scientists, IT, and business teams for thorough evaluation, testing, and validation of ML models.
- **Code Sharing**: Promote open code sharing to enhance transparency and foster discussions on best practices, avoiding overly complex algorithms.
- **Transparent Decision-Making**: Clearly explain the rationale behind ML decisions and parameter choices to improve understanding and accountability.
- **Uncertainty Sets**: Use uncertainty sets like conformal prediction to reduce risks and improve confidence in model outputs.

# F. IT issues (ML infrastructure)

Are the IT prerequisites for ML solutions in place?

- **Outdated IT systems**: Many NSOs lack the infrastructure needed for ML innovation.
- **Resource challenges:** Insufficient resources and unclear governance slow innovation.
- **Complex production integration:** Significant effort required to integrate ML into production.
- **IT overload:** IT teams are stretched between modernisation and supporting new tech.
- **Cloud adaptation issues:** Customising cloud environments for NSO needs is demanding.
- **Skill shortages:** NSOs struggle with acquiring and developing ML and cloud skills.
- **Vendor lock-in:** Dependence on specific providers limits flexibility and future transitions.

# Thanks for listening. Questions?

Steffen Moritz
steffen.moritz@destatis.de