EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Quality Dimensions of Machine Learning in Official Statistics

Younes Saidani

joint work with: Florian Dumpert, Christian Borgs, Alexander Brand, Andreas Nickl, Alexandra Rittmann, Johannes Rohde, Christian Salwiczek, Nina Storfinger, Selina Straub

Federal Statistical Office Germany

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union

Motivation

- Quality frameworks: Code of Practice & Quality Assurance Framework
- CoP & QAF require official statistics to be "constantly striving for innovation", but are derived from the needs of "traditional" statistics production
- "Innovative" statistical methods can differ substantially from traditional ones:
    - Certain quality dimensions may not be applicable to new methods at all
    - They may be applicable in principle but differ with regards to methodological details
    - New methods may present new challenges that are not covered by existing dimensions

→ Need to **assess the compatibility** of new methods with existing official statistics quality frameworks, and to **offer accompanying quality guidance** when required

How useful is existing quality guidance for ML?

- Many quality principles in the CoP & quality indicators in the QAF are potentially affected by ML methods or can be used to derive quality requirements for their usage (principles for statistical processes and statistical outputs: 7.1-7.7, 8.3-8.5, 9.1, 9.6, 10.2, 10.4, 12.1, 12.2, 13.1, 13.5, 14.2, 15.1, 15.5)
- On the one hand: broad enough to cover methodological characteristics of ML

→ Quality guidance for ML should **build on existing quality frameworks**

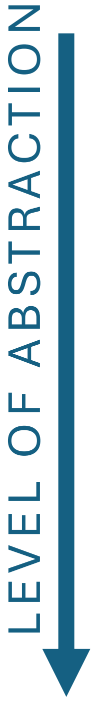
- On the other hand: not specific enough to provide useful guidance in practice

→ Need for developing **quality guidance specifically tailored** to ML

Our approach

- Multi-step process (abstract to specific), analogous to CoP & QAF: quality principles -> quality indicators -> quality methods
- What is needed to ensure compatibility of ML applications used in official statistics production with existing official statistics quality standards?
  1. **Quality dimensions**: What does it mean for ML to have „high quality“?
  2. **Quality guidelines**: How to implement quality along the above dimensions during development?
  3. **Quality indicators & metrics**: How to evaluate quality in development & production?
  4. **Quality documentation**: How to communicate quality of ML in an appropriate, standardised way?
- Work-in-progress: 1. completed, 2. in progress, 3. & 4. pending

Proposed quality dimensions



• Accuracy	*Predictions*	„phenomenon is described correctly“
• **Robustness**	*Model*	„stable results despite small perturbations“
• **Explainability**		„understand how results are generated“
• Reproducibility	*IT infrastructure*	„reproduce identical results“
• Timeliness & Punctuality	*Business processes*	„deliver up-to-date results punctually“
• Cost-effectiveness		„appropriate costs“

+ Cross-cutting issues: **MLOps**, **Fairness**

Robustness

- The degree to which a model produces stable (but useful) results given small perturbations in the environment
- Possible perturbations: outliers in the data, changes in its distribution, violations of model assumptions, structural changes in the observed phenomenon over time (concept drift), or different choices of hyperparameters
- Which “results” should be stable? Plausible candidates:
  - specific predictions (e.g. for influential data points)
  - model coefficients
  - accuracy metrics
  - aggregates that are produced downstream in the statistical production process (e.g. total revenue by industry, export volume by enterprise type)

Fairness

- Aim: to avoid treating certain groups unjustifiably differently in a relevant way by or as a result of statistical procedures (like ML)
- In the context of official statistics, such effects are usually indirect, e.g., through political decisions based on the published data
- Example: statistical aggregates are systematically over- or underestimated for certain sub-groups (e.g., economic sectors, types of households, regions, ...)
- Connections to accuracy (imbalanced data) and explainability

MLOps

- Aim: best possible fulfilment of the quality dimensions
- Necessary: Establishment of standardised processes for data processing, data management and model maintenance
- Strong connections to reproducibility, but also to explainability, timeliness & punctuality and cost-effectiveness

(Of course, we were not the first that stated this for official statistics; see, e.g., Engdahl J, Choi I, Deeben E, Karanka J, Karlsson A, Meszaros M, Pocknee J, Holroyd P, Baily A (2022) Building an ML Ecosystem in Statistical Organisations)

Lessons learned

- Task at hand requires collaboration of ML practitioners, subject-matter statisticians and quality officers
- Must strike a balance between being overly general (and thus not useful) and being too specific (and thus only applicable to certain ML methods)
- Must consider that ML best practices are rapidly evolving
- Theoretical musings are only useful if implemented in practice, ensuring adoption of new standards is of utmost importance!

Thank you

Younes Saidani

Data Scientist

Federal Statistical Office Germany

younes.saidani@destatis.de

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union