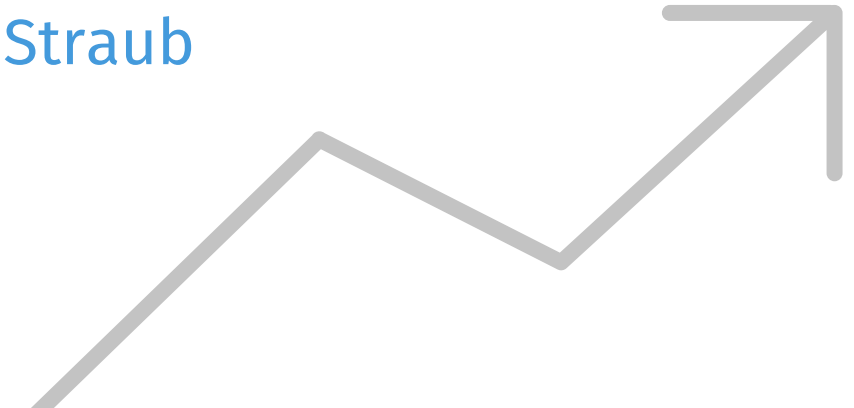
75 DESTATIS Statistisches Bundesamt

A Quality Concept for the Use of Machine Learning in Official Statistics

Florian Dumpert

joint work with Younes Saidani, Christian Borgs, Alexander Brand, Andreas Nickl, Alexandra Rittmann, Johannes Rohde, Christian Salwiczek, Nina Storfinger and Selina Straub

Saidani Y, Dumpert F, Borgs C, Brand A, Nickl A, Rittmann A, Rohde J, Salwiczek C, Storfinger N, Straub S (2023) Qualitätsdimensionen Maschinellen Lernens in der Amtlichen Statistik. Submitted to AStA Wirtschafts- und Sozialstatistisches Archiv



Bad quality reduces trust very, very fast.

Walter Radermacher, 2022

Starting Points

- **Quality Framework for Statistical Algorithms**[1)] with its dimensions
  - Accuracy
  - Explainability
  - Reproducibility
  - Timeliness
  - Cost-effectiveness
- **Quality frameworks**[2)]

1) Yung et al (2022) A quality framework for statistical algorithms. Statistical Journal of the IAOS, 38(1), 291–308
2) https://www.destatis.de/DE/Methoden/Qualitaet/qualitaetshandbuch.html, https://ec.europa.eu/eurostat/web/quality/european-quality-standards/quality-assurance-framework, https://unstats.un.org/unsd/unsystem/documents/UNSQAF-2018.pdf

Our contribution

(1) Identification of relevant quality dimensions for ML by **analysing the quality principles** contained in the European Statistics Code of Practice

(2) in light of the **methodological peculiarities** of ML

whereas

(2a) **robustness** is proposed as a stand-alone quality dimension

(2b) machine learning operations (**MLOps**) and **fairness** are discussed as two cross-cutting issues

(2c) suggestions are made how quality assurance can be **conducted in practice** for each quality dimension.

Quality dimensions

Modell

- Accuracy
- Robustness
- Explainability

Infrastructure

- Reproducibility

Processes

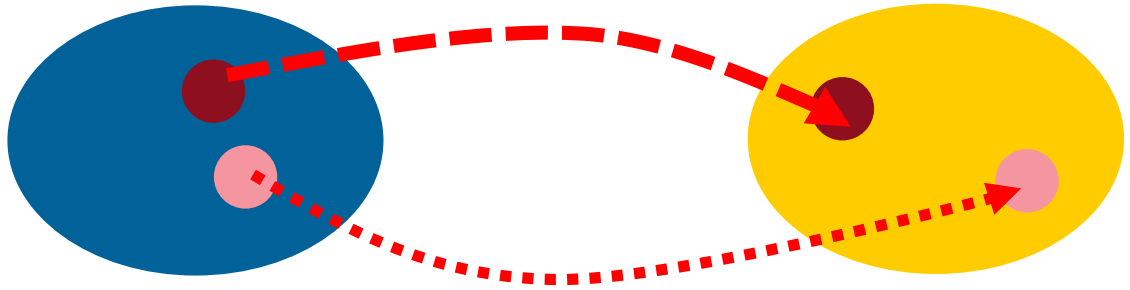
- Timeliness & Punctuality
- Cost-effectiveness

+ Fairness
+ MLOps

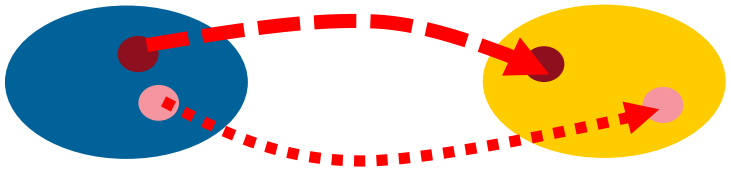
Robustness

General Aspects

- Definition:
  The property of a model to give stable results when the data or the model is slightly perturbed.
- ≠ Accuracy (sample-related confidence or prediction intervals)
- ≠ Replicability (in a scientific sense)



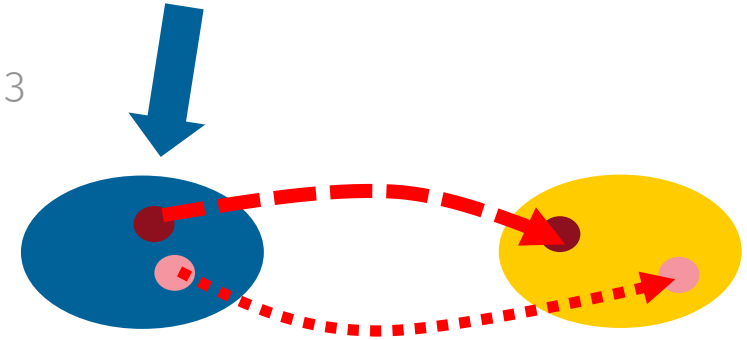
Robustness



Examples

- $f: \mathbb{X} \to \mathbb{Y}, f(x) = 0 \ \ \forall\, x \in \mathbb{X}$ is highly robust – but useless
- $f: \mathbb{X} \to \mathbb{Y}, f(x) = f(x_1, \dots, x_n) = median(x_1, \dots, x_n)$ is more robust than $f: \mathbb{X} \to \mathbb{Y}, f(x) = f(x_1, \dots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Simpler models might be more robust than highly complex models
- Overfitted models might be less robust than models that generalise well

Robustness



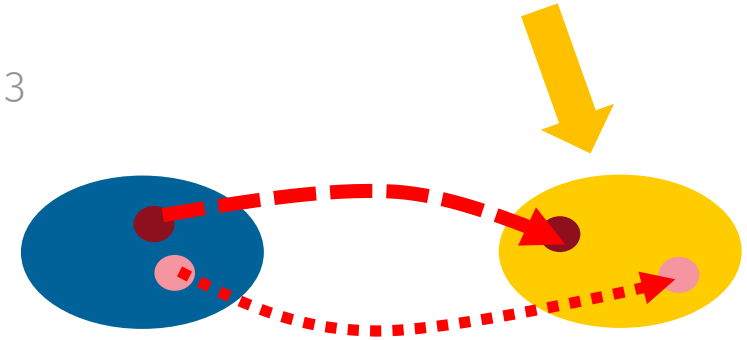
Possible Perturbations

- (Different **samples** from the same population)
- **Noise** on some observations
- **Errors** (including typos, including intentional changes) in some observations

Special situation that needs special treatment:

- Concept drift (i.e., changes in the distribution of input variables, of the target variable(s) or of the relationship between them)

Robustness



Possible Targets

- Every single prediction
- Performance measures for the ML algorithm (e.g., F1 measure)
- Coefficients/parameter of the ML algorithm (e.g., weights of a DNN, $\beta_0, \ldots, \beta_d$ in a parametric model)
- (Final) statistical output of interest (i.e., downstream quantities; e.g., median turnover per economic sector)

→ To discuss with subject matter experts … and **depending on how integrated processes and data management are**

Robustness

Proposals for Quality Guidelines

1. Target variables that are to be the subject of robustness have been defined.
2. Robust procedures were considered in the pre-selection of ML procedures to be investigated.
3. The effects of data perturbations on the target variable in the form of a suitable resampling procedure were investigated.
4. The effects of data perturbations in the form of deliberate manipulation or addition of grossly incorrect data points were investigated.
5. The effects of model perturbations in the form of violating assumptions about the model were investigated.
6. A procedure for the detection of concept drift was implemented.
7. The effects of model perturbations in the form of different hyperparameter choices were investigated.

Fairness and MLOps

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

Fairness and MLOps

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 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)

Summary

- **No quality, no trust**
- **Quality dimensions** for the use of machine learning in official statistics were proposed on the basis of preliminary work within the QF4SA
- Some first proposals for **quality guidelines** were presented
  (in the paper for all dimensions; for robustness during the talk)

Contact

Statistisches Bundesamt
65180 Wiesbaden
Germany

Florian Dumpert
florian.dumpert@destatis.de
Phone +49 611 75-3887

www.destatis.de

www.destatis.de/kontakt