Quality dimensions of machine learning in official statistics

Younes Saidani[1], Florian Dumpert[1]

[1]Federal Statistical Office of Germany

Abstract

Official statistics distinguishes itself through the legally stipulated requirement to ensure the quality of its publications. To this end, it adheres to European quality guidance, which is operationalised at the national level in the form of quality manuals. Hitherto, these have been designed and interpreted with the requirements of "classical" statistical production processes in mind. Thus, in order to ensure continued adherence to quality standards, tailored quality guidance must be developed to accompany the increasing use of machine learning (ML) methods in official statistics. This paper sets out a multi-step approach towards achieving such guidance for ML, builds on previous work to suggest six quality dimensions for ML, and discusses fairness and machine learning operations (MLOps) as two cross-cutting issues with relevance to most quality dimensions.

Keywords: machine learning, explainability, interpretability, robustness, stability

1. The need for tailor-made quality guidance for ML

Official statistics enjoys special privileges over other statistics providers: its work is not subject to market forces and profitability concerns, and respondents are in many cases legally required to respond to its surveys, thus enabling statistical offices to continuously collect and publish data on topics of public interest. In turn, it is expected that official statistics produces data of high quality in order to fulfil its important role as a reliable data provider. After all, "bad quality erodes reduces trust [in official statistics] very, very fast", as Walter Radermacher, former President of the German Federal Statistical Office and Director-General of Eurostat, once remarked. The quality of statistical data has therefore always been of great importance in official statistics.

But what does it mean for data to be of "high quality"? In Europe, the Statistics Code of Practice (CoP) defines 16 key principles for the institutional environment, statistical processes and statistical outputs, which are used to assess and safeguard quality in statistical offices. The Quality Assurance Framework (QAF) provides further guidance on how to implement the high-level principles from the CoP by offering more detailed methods, tools and good practices.

Among other things, the CoP and QAF explicitly require official statistics to be "constantly striving for innovation" (QAF, Indicator 7.1) to improve the quality of its products. Yet both documents are derived from and based on the requirements and challenges of "traditional" statistics production. This poses a practical challenge, since innovative statistical methods can differ

substantially from traditional ones, potentially reducing the usefulness of existing quality frameworks in one of three ways:

1. Certain quality dimensions may not be applicable to new methods at all,
2. they may be applicable in principle but differ with regards to the methodological details,
3. or new methods may present new challenges that are not covered by existing quality dimensions.

Consequently, upon adopting new methods, there is a need to assess their compatibility with existing official statistics quality frameworks, and to offer accompanying quality guidance in case it is needed.

Machine learning (ML) – an example of such an "innovation" – has recently matured from future technology to industry standard. The term refers to a collection of methods that (according to one definition) differ from traditional statistical methods with regards to their intended use: While "classical", research-oriented statistics generally focuses on hypothesis testing, ML algorithms mostly aim to optimally predict the properties of new observations, often with the aim of process automation. ML methods such as tree-based approaches (including random forests and boosting), support vector machines and neural networks offer great potential for classification and coding tasks, error detection and correction as well as the imputation of missing values. As a result, they have been actively taken up and piloted in official statistics (e.g. in Germany, see Dumpert 2023). Yet the adoption of ML is marred with methodological, technological and regulatory challenges – among them the question if and how existing quality guidance from official statistics can be applied to ML methods.

A closer look reveals that a number of quality principles contained in the CoP and many of their respective quality indicators from the CAF are either potentially affected by machine learning methods or can be used to derive quality requirements for their usage – in particular, the quality principles for statistical processes and statistical outputs. Relevant quality indicators include 7.1 to 7.7, 8.3 to 8.5, 9.1 and 9.6, 10.2 and 10.4, 12.1 and 12.2, 13.1 and 13.5, 14.2 and 14.2, and 15.1 and 15.5. Elaborating on each quality indicator and its relevance to ML is beyond the scope of this paper, but suffice to say that they are broad enough to cover most methodological characteristics of ML. Yet they are also not specific enough to provide useful guidance for the use of ML methods in official statistics: the principles for statistical processes (7 to 10) are too abstract, and the principles for statistical products (11 to 15) – referring to the quality of statistical publications rather than that of intermediate results, the generation of which is where ML methods are most often employed – are too indirect. Discussing the potential need for new quality criteria when integrating new data and methods in official statistics, Broe

et al. (2021) similarly conclude that “the methodological quality aspects related to statistical learning and big data clearly contain new elements compared to the well-established ESS [European Statistical System] output quality dimensions. Those new elements can be seen as extensions of the well-established dimensions rather than completely new quality dimensions of their own” (ibid., p. 357). On the one hand, this underscores the need for developing quality guidance specifically tailored to ML in order to enable adequate quality management (see also Julien 2020 and Dumpert 2021). On the other hand, it shows that such new guidance can – and in fact should – build on the existing quality frameworks for official statistics, highlighting relevant links where possible, providing further details where useful, and offering extensions where needed.

This paper aims to contribute to the development of such new quality guidance for machine learning methods in official statistics. It is based on previous, collaborative work in German by Saidani et al. (2023).

2. A four-step approach towards comprehensive quality guidance for ML

Attempting to establish an extension of existing quality frameworks for machine learning methods is a multi-step process. Similar to existing quality guidance – which starts with high-level quality principles (in the CoP) that are then broken down into quality indicators and further supplemented by so-called methods (in the QAF) – the following four-step structure is suggested for ML, starting at the abstract level and then moving into specifics:

1. **Quality dimensions** that define what “quality” means for ML, i.e. what is meant by the concept and what are its primary components.
2. **Quality guidelines** for statistical processes that describe what needs to be considered during ML development in order to ensure quality along the above dimensions.
3. **Quality indicators and metrics** for statistical results generated using ML that permit a quantitative evaluation of quality along the above dimensions during development and production.
4. Standards and recommendations for **quality documentation** that aid in communicating the quality of ML used in statistics productions in an appropriate, standardised manner.

All four steps are necessary components of a holistic quality guidance for ML in official statistics. Quality dimensions provide the conceptual background, quality guidelines guide the design of processes and quality indicators formulate adequate metrics for evaluating quality. Last but not least, given the regulatory background and increasing user requirements, standardised

quality documentation ensures transparency about the ML methods used and their effect on the quality of statistical products.

Developing such dimensions, guidelines, metrics and documentations requires the collaboration of machine learning practitioners, subject-matter statisticians and quality officers. Given the plethora of methods that can be considered “machine learning”, they must strike a balance between being overly general – and thus not useful – and too specific – and thus only applicable to certain methods. They must also consider that ML is a rapidly evolving field and thus allow for changing best practices. Last but not least, theoretical musings are only useful if they are implemented in practice; thus ensuring adoption of new standards in statistical offices – the difficulty of which must not be underestimated due to cultural and behavioural obstacles – is of utmost importance.

Besides structuring the task at hand, this paper contributes by formulating high-level quality dimensions tailored to the methodological peculiarities of ML, thereby expanding on previous suggestions, and by introducing two cross-cutting issues that are highly relevant to the quality of ML algorithms. Work on quality guidelines based on these dimensions is currently ongoing. Subsequent work should aim to devise quality metrics and standardised quality documentations.

3. Quality dimensions

As part of the evaluation of ML methods for use in official statistics, colleagues from statistical offices have already attended to the task of deriving a set of suitable quality dimensions. Building on Broe et al. (2021) – perhaps the first such contribution – a group of national experts from UNECE member states and Australia under the umbrella of the UNECE High-Level Group for the Modernisation of Official Statistics developed a “Quality Framework for Statistical Algorithms” (Yung et al. 2022). In it, they take up many of the quality aspects elaborated by Broe et al., further specify them with regard to the methodological peculiarities of “statistical algorithms” (including ML algorithms) in statistics production, and suggest five quality dimensions: explainability, accuracy, reproducibility, timeliness and cost effectiveness.

While these dimensions cover many of challenges and advantages associated with ML, they omit at least one important topic: How reliable is a model, once trained, in production, when it is confronted with previously unanticipated conditions? For instance, relationships that a model has learned from the training data may change in the real world over time. This and related problems are covered by the term “robustness” (also: “stability”), which is essential for an adequate quality assessment of machine learning methods.

Once robustness is added to the above list, the six quality dimensions cover all indicators from the QAF that are relevant for the use of ML: Indicators associated with sound methodology, appropriate statistical procedures and non-excessive burden on respondents can be assigned to the dimensions accuracy, robustness, explainability and reproducibility. Relevant indicators from accuracy and reliability are split up into the two dimensions accuracy and robustness, the latter of which also covers coherence and comparability. Accessibility and clarity relate to the dimensions reproducibility and explainability. The principles timeliness and punctuality as well as cost-effectiveness remain as dimensions with the same name.

Table 1: Quality dimensions for machine learning summarised

Dimension	Description	Level of abstraction
Accuracy	Phenomenon is described correctly	Predictions
Robustness	Stable results despite small perturbations	Predictions, Model
Explainability	Understand how results are generated	Predictions, Model
Reproducibility	Reproduce results identically	IT infrastructure
Timeliness and punctuality	Deliver up-to-date results punctually	IT infrastructure, Business processes
Cost-effectiveness	Appropriate costs	Business processes

Table 1 summarises the six quality dimensions and offers a brief description. Furthermore, it arranges the dimensions in increasing order of abstraction: While accuracy and (partly) robustness concern themselves with individual predictions, cost-effectiveness is assessed at the level of business processes. In the following, each quality dimension is briefly discussed in more detail.

3.1 Accuracy

Accuracy is the degree to which a statistical output is able to correctly describe the phenomenon being measured, i.e. to minimise the suitably measured distance between the estimate and the true value. Accuracy is thus not a binary criterion; how much accuracy suffices depends on the specific use case. The metrics used to evaluate accuracy depend on the phenomenon under observation, in particular whether the goal is classification or regression. Which metric is deemed most relevant for model selection depends, again, on the use case and is a decision to be made by subject matter experts. In any case, accuracy should not just be reported as a point estimate, but also accompanied by a measure of uncertainty such as a

confidence interval. Furthermore, model accuracy should also consider the uncertainty or bias in the training, validation and test datasets. Thus, ensuring high-quality training data is essential for accurate ML models.

3.2 Robustness

Robustness is the degree to which a model produces stable (but useful) results given small perturbations in the environment – which may be 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. Concept drift in particular is almost certainly an issue if a model is employed for multiple years, and can be dealt with by regular retraining or by implementing a mechanism for drift detection.

But which “results” should be stable given small perturbations? Plausible candidates include specific predictions (e.g. for influential data points), model coefficients, accuracy metrics or aggregates that are produced downstream in the statistical production process (e.g. total revenue by industry, export volume by enterprise type). The latter seems most relevant in almost all cases, yet the large number of processing steps – often conducted using separate tools that may or may not be connected by automated interfaces – makes assessing robustness on this level very difficult. In practice, the most feasible and useful approach is to evaluate the stability of accuracy metrics under simulated, adverse scenarios.

3.3 Explainability

Explainability is the ability to understand which relationships the algorithm uses to make predictions, i.e. to be able to demonstrate the (possibly local) relationship between input and output variables. Defined in this way, explainability as post-hoc interpretability is becoming more and more relevant in the face of new laws and regulations on the use of artificial intelligence systems (European AI Act). Independently thereof, explainable models are preferable because they generally increase trust among users and allow developers to spot specification mistakes more easily.

3.4 Reproducibility

Reproducibility is the ability to achieve identical results when using the same data, the same algorithm and the same computing environment. In practice, this is achieved by versioning, documenting and archiving data, codes and libraries.

3.5 Timeliness and punctuality

Timeliness and punctuality describes the ability to design, train and apply the ML algorithm within the required time frame, and to publish up-to-date results.

3.6 Cost-effectiveness

Cost-effectiveness describes the implementation costs relative to the abovementioned quality dimensions. Thus, lower costs given fixed accuracy, robustness, explainability, reproducibility, timeliness and punctuality imply better cost effectiveness.

4. Cross-cutting issues

Last but not least, there are two cross-cutting topics that are not quality dimensions themselves, but are nonetheless closely related to the six dimensions discussed above: fairness – whether an ML model produces "fair" results for subgroups of interest – and MLOps – a set of technical standards, which in practice are necessary for a time- and cost-efficient implementation of many quality requirements (especially those associated with reproducibility).

4.1 Fairness

The fairness of statistical procedures refers to the effects that algorithmic decisions or classifications can have on the individuals or administrative units surveyed. In the context of official statistics, such effects are usually indirect, for example through political decisions based on the published data. In addition to general considerations such as the quality of the training data – a machine learning model can learn false correlations regardless of the estimation accuracy if the training data already contain structural biases (Mehrabi et al. 2022) – the accuracy of an ML procedure can have implications for fairness if statistical aggregates for certain subgroups are systematically over- or underestimated. This is particularly relevant for subgroups that are less represented in the data or tend to be at the edge of the distribution. Strategies to increase the accuracy of the model for small subgroups and thus avoid bias include up- and downsampling methods, hybrid forms such as SMOTE and ROSE, active learning using expert feedback, weighting observations, or the use of adapted ML models that are specially optimised for unbalanced data.

In practice, these methods are not absolutely necessary if simpler models provide satisfactory estimation accuracy for subgroups. This can be evaluated by calculating whether simple aggregates for subgroups of interest are underestimated or overestimated (to a relevant extent) in the trained model.

Besides accuracy, fairness also relates to explainability in so far as understanding the effect of subgroup characteristics on the target variable allows ML developers to draw valuable conclusions about the inner workings and reliability of the model. For instance, a model that has learned multivariate relationships that are supported by empirical research or common-sense might inspire confidence in its ability to deal with unseen data.

4.2 MLOps

Standardised data processing and data management are essential for quality: The evaluation of accuracy is facilitated immensely by tools that output relevant quality metrics and their variance during model training by default. Ensuring robustness and explainability can be extremely time-consuming if specific tests and evaluation routines are not pre-programmed and easily accessible. Standardised processes, once developed, also allow for better timeliness and higher efficiency. In order to ensure that data, codes and environments are reproducible, certain procedures and technical tools must be established and used across the whole organisation.

Practices and processes that aim to reliably and efficiently develop, productively deploy, manage, monitor and maintain machine learning models are summarised under the term “machine learning operations” (MLOps) (Kreuzberger et al. 2022). To the extent that such processes, tools and practices are necessary to fulfil the quality dimensions specified in the quality frameworks of official statistics, MLOps is also a basic prerequisite for machine learning in official statistics. Consequently, it is essential that MLOps is given ample consideration during the development of IT platforms and data management systems. As part of the ONS-UNECE-ML-2022 project, requirements for MLOps systems and possible system architectures have been gathered and evaluated (Engdahl et al. 2022).

5. Conclusion

Since ML methods differ considerably from conventional statistical methods, existing quality frameworks cannot be applied without further specification. This paper has aimed to contribute to the development of a tailor-mode, comprehensive guidance for the use of ML, thus paving the way for the widespread usage of such methods in official statistics. Further work is required on quality guidelines, quality indicators and standards for quality documentation. Subsequently, theoretical standards need to be implemented in day-to-day statistics production.

Acknowledgment

This paper is based on previous work in German, conducted by the author in collaboration with Florian Dumpert, Christian Borgs, Alexander Brand, Andreas Nickl, Alexandra Rittmann, Johannes Rohde, Christian Salwiczek, Nina Storfinger and Selina Straub, and published as Saidani et al. (2023).

References

De Broe, S., Struijs, P., Daas, P., Van Delden, A., Burger, J., Van Brakel, J., Bosch, O. T., Zeelenberg, K. & Ypma, W. F. H. (2021). Updating the paradigm of official statistics: New quality criteria for integrating new data and methods in official statistics. *Statistical Journal Of The IAOS*, 37(1), 343–360. https://doi.org/10.3233/sji-200711

Dumpert, F. (2023). Machine Learning in German Official Statistics. In: Snijkers, G., Bavdaž, M., Bender, S., Jones, J., MacFeely, S., Sakshaug, J.W., Thompson, K.J. & van Delden, A.: *Advances in Business Statistics, Methods and Data Collection*, 537–560.

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. URL: www.statswiki.unece.org/display/ML/Machine+Learning+Group+2022

Julien, C. (2020). UNECE – HLG-MOS Machine Learning Project. URL: www.statswiki.unece.org/display/ML/Machine+Learning+Project+Report

Kreuzberger, D., Kühl, N. & Hirschl, S. (2022). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.02302

Mehrabi, N., Morstatter, F., Saxena, N. A., Lerman, K. & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35. https://doi.org/10.1145/3457607

Saidani, Y., Dumpert, F., Borgs, C., Brand, A., Nickl, A., Rittmann, A., Rohde, J., Salwiczek, C., Storfinger, N. & Strauß, S. (2023). Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik. *AStA. Wirtschafts- und Sozialstatistisches Archiv*, 17(3–4), 253–303. https://doi.org/10.1007/s11943-023-00329-7

Yung, W., Tam, S. M., Buelens, B., Chipman, H. A., Dumpert, F., Ascari, G., Rocci, F., Burger, J. & Choi, I. (2022). A quality framework for statistical algorithms. *Statistical Journal Of The IAOS*, 38(1), 291–308. https://doi.org/10.3233/sji-210875