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

**Florian Dumpert** (Federal Statistical Office of Germany)
(joint work with Younes Saidani, Christian Borgs, Alexander Brand, Andreas Nickl, Alexandra Rittmann, Johannes Rohde, Christian Salwiczek, Nina Storfinger and Selina Straub)

## A) Background

1. For many years, official statistics have been concerned with questions of quality. This is imperative, since on the one hand it plays an important role in decisions of parliament, government, jurisdiction, economy and society, but on the other hand it also lives and depends on the trust of these addressees as well as of the respondents. Quality is therefore a necessary criterion for the existence, good work and high acceptance of official statistics. This statement is valid for the statistical institutions of many states and for supranational and international statistics-producing institutions. Therefore, many of them have introduced quality frameworks in which quality dimensions are defined. Among others, the frameworks of Canada (Statistics Canada 2019), Australia (ABS 2009), the European Union (Eurostat 2017) or the United Nations (UN 2018) can be mentioned here.

2. In the past ten years, machine learning has increasingly found its way into official statistics (e.g., Beck et al 2018). In this context, machine learning should be described less as a class of functions, but rather – in the sense of Breiman (2001) – via the goal of using methods and procedures. We speak of (statistical) machine learning when not hypothesis tests and causal effects are of primary interest, but rather the predictive power of the procedures. The more the predictive aspect is in the foreground of an application (and less the inference or the deeper explanatory power), the more we attribute a procedure to machine learning. Examples in this sense are often (but not exclusively) the use of deep neural networks (deep learning), support vector machines, boosting methods and random forests. In fact, these are always semi-parametric and non-parametric statistical methods, in particular these are methods that recognise patterns on a data-driven basis (i.e. learn them during training); the fewer assumptions are made about the functional relationship between the explanatory variables and the variable to be explained (target variable), the more non-parametric a method is and the higher the number of cases of observations that are usually required to learn the patterns and relationships.

3. These methods are thus finding their way into official statistics, very often in the form of supervised learning. Supervised learning means that data sets are available that contain explanatory variables and the corresponding values of the target variables, i.e. that a kind of truth (also referred to as ground truth) is presented to the procedure during training. How well the ground truth reflects real associations during training is often impossible to judge. Nevertheless, it is important to ensure that the ground truth presented, i.e. the training data set used, is correct and sufficiently comprehensive according to a consensus of experts.

4. With regard to the purpose of use, it can be seen that machine learning is often (although not only) used in phases 4 and 5 of the Generic Statistical Business Process Model (GSBPM; UNECE 2019), i.e. to generate intermediate results rather than statistical end products. Phase 4 involves the collection of data (collect), while phase 5 involves the processing of the data (process). The use of machine learning often serves to (partially) automate production steps. Unlike rule-based (and thus deterministic) automation, machine learning works data-based (and thus rather empirically or inductively). This may sound dangerous from a quality point of view, but it does offer advantages: On the one hand, a large or even complex system of deterministic rules during e.g. classification and coding of data is not easy to write down. A machine learning procedure learns such a system explicitly or implicitly by training on the basis of given data. On the other hand a machine learning approach can react to changes in rules based on data; hard-coded rules would require manual rework.

5. The existing quality frameworks already contain specifications for the individual phases of GSBPM. They are, as a quality framework should be, designed on a high level of abstraction. For example, they

contain specifications for classification and coding in general, but no special regulations for concrete situations and thus no regulations for the classification or coding to be carried out by a machine learning procedure. To stay in this example, it could be, for example, answering the question whether an observation is actually relevant in the sense of the definition underlying the statistic. Another example would be the assignment of observations to classes of a classification scheme (occupations, economic activity, consumer goods, etc.). It is clear that machine learning must fulfil these abstract requirements and can partly also contribute to an improvement in the fulfilment of requirements (e.g. to timeliness through faster production by automating sub-processes). However, it is still unclear how these requirements can be concretised for the use of machine learning.

6. Workpackage 2 of the UNECE HLG-MOS Machine Learning Project 2019/2020 therefore initially dealt with the question of how to reconcile machine learning and quality in official statistics. The result was the QF4SA (Quality Framework for Statistical Algorithms; Yung et al 2022), which represents a first draft to concretise the above-mentioned, but so far missing, aspects. Basic principles for the use of statistical machine learning in official statistics are discussed on the basis of five dimensions: Accuracy, explainnability, reproducibility, timeliness, cost effectiveness. The QF4SA is explicitly "a first attempt to lay down some groundwork to guide official statisticians in comparing algorithms (be they traditional or modern) in producing official statistics." (Yung et al 2022, p. 306).

7. Building on the QF4SA, German official statistics have begun to consider further and to expand and further differentiate the QF4SA.

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

8. The analysis of the QF4SA showed that further conceptual work is necessary in order to be able to safely guarantee the use of machine learning in official statistics also from the point of view of the quality frameworks. On the one hand, a further quality dimension (robustness) as well as further secondary aspects (fairness and MLOps) were explicitly introduced; on the other hand, the dimensions already mentioned by the QF4SA were examined and discussed in greater depth. Furthermore, the work already contains initial proposals for operationalising the quality dimensions in the concrete work in the statistical offices.

9. The requirements and quality dimensions were explicitly derived from the quality dimensions already existing in the general frameworks for the processes and products of official statistics. On the one hand, this was done in order not to create the quality dimensions for the use of machine learning arbitrarily, but to link them to the existing frameworks, but on the other hand also to emphasise the positive effects of taking the quality dimensions into account with regard to general statistical production.

10. The work identifies the following six dimensions of quality (Saidani et al 2023):

(i) Accuracy: The degree to which a statistical output correctly describes the phenomenon being measured. Exemplary criteria for assessing the fulfilment of the quality dimension: Relevant quality measures were identified, point estimators and confidence intervals were given, accuracy of estimation for subgroups was evaluated.

(ii) Robustness: The property of a machine learning model to produce stable results in the presence of small pertubations of the data or the model. Example criteria: Effects of data perturbations were investigated, effects of model perturbations were investigated, procedure for detecting possible concept drifts was implemented.

(iii) Explainability: The ability to understand what relationships the algorithm uses to make predictions. Example criteria: Requirements for explainability have been determined, relationship between input and output variables is explainable.

(iv) Reproducibility: The ability to obtain identical results using the same data and algorithm. Example criteria: Data used was frozen and archived, code was versioned and archived, code and data were documented, software used and versions were documented.

(v) Timeliness and punctuality: The ability to design, train and apply the algorithm within the required time and to publish up-to-date results. Example criteria: Sufficient time has been planned for data acquisition, sufficient time has been planned for conception, selection and testing of possible ML methods, final results can be published earlier than before while maintaining the same quality.

(vi) Cost-effectiveness: The ratio of the extent of the other quality dimensions to the costs of implementation. Possible expenses: (a) one-time: development of ML procedures, concepts, models and algorithms, creation or procurement of training and test data, transfer to standardised procedures and development of standard tools; (b) ongoing: expenses for cleaning and preparing training and test data as well as for labelling data in classifications, testing of new technologies and procedures in the field of ML, maintenance, care and updating of models and algorithms.

11. The reason why robustness was included as an additional quality dimension is that this dimension was hinted at in the QF4SA, but not elaborated to the necessary extent. However, it is crucial for the valid use of machine learning in official statistics. If, for example, the sub-steps automated by machine learning were to make completely different estimates or assignments for similar input values, subject matter experts and the public would no longer have confidence in this sub-step. Inclusion as a separate quality dimension therefore seemed necessary.

**Final remark:** The considerations presented here were written down in detail in a paper by colleagues from German official statistics (Saidani et al 2023), which is currently under review.

**References:**

ABS (Australian Bureau of Statistics) (2009) The ABS Data Quality Framework. https://www.abs.gov.au/websitedbs/D3310114.nsf//home/Quality:+The+ABS+Data+Quality+Framework

Beck M, Dumpert F, Feuerhake J (2018) Machine Learning in Official Statistics. https://doi.org/10.48550/arXiv.1812.10422

Breiman L (2001) Statistical modeling: The two cultures. Statistical Science, 16(3), 199–231

Eurostat (2017) European Statistics Code of Practice. https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice

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

Statistics Canada (2019) Quality Assurance Framework, 3rd edition. https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm

UNECE (United Nations Economic Commission for Europe) (2019) Generic Statistical Business Process Model (GSBPM). https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1

United Nations (2018) UN Statistics Quality Assurance Framework Including a Generic Statistical Quality Assurance Framework for a UN Agency. https://unstats.un.org/unsd/unsystem/documents/UNSQAF-2018.pdf

Yung W, Tam S-M, Buelens B, Chipman H, 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