Quality Framework for Statistical Algorithm (QF4SA)

InKyung Choi

UN Economic Commission for Europe

Statistics Division

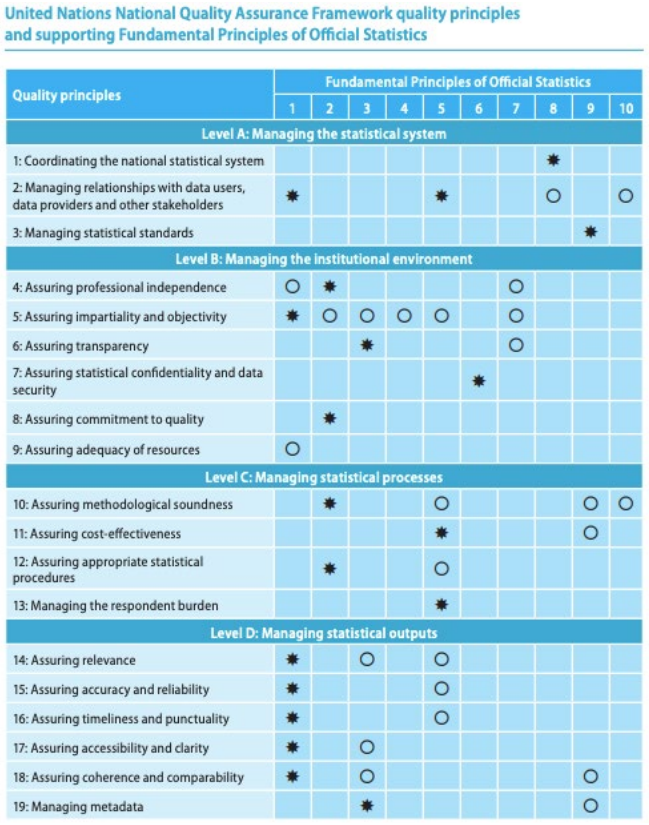
Machine Learning for Official Statistics Workshop (June 6, 2023)

Content

1. Background
2. Quality dimensions in QF4SA
    1) Accuracy
    2) Explainability
    3) Reproducibility
    4) Timeliness
    5) Cost-effectiveness
3. Summary

1. Background

- National Statistical Offices (NSOs) and international statistical organisations are the provider of official statistics and have a responsibility to ensure that the highest quality outputs are produced
- Quality frameworks to support quality assurance
- With increasing interest in machine learning methods, existing quality frameworks need to be looked at



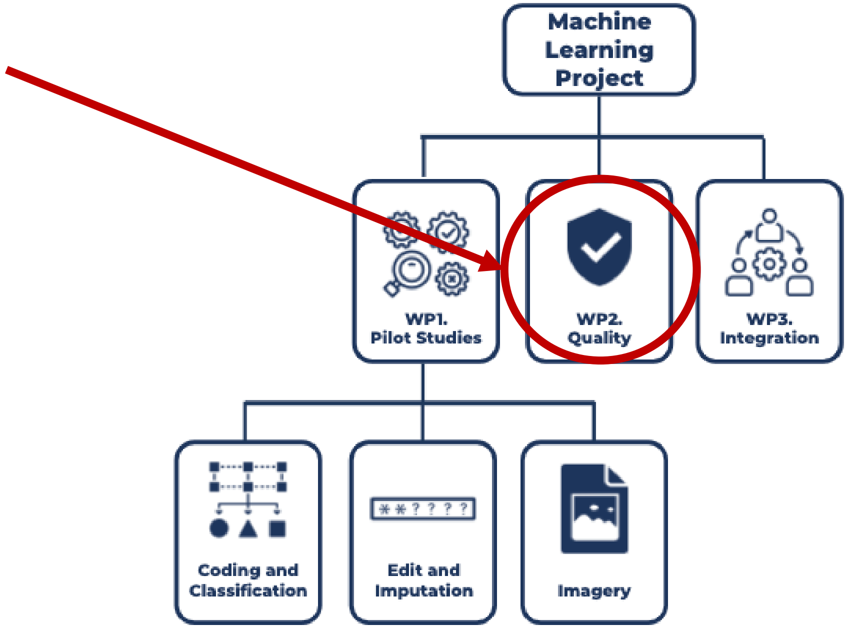
United Nations National Quality Assurance Framework quality principles and supporting Fundamental Principles of Official Statistics

Quality principles	Fundamental Principles of Official Statistics									
	1	2	3	4	5	6	7	8	9	10
Level A: Managing the statistical system										
1: Coordinating the national statistical system								✸		
2: Managing relationships with data users, data providers and other stakeholders	✸				✸			○		○
3: Managing statistical standards									✸	
Level B: Managing the institutional environment										
4: Assuring professional independence	○	✸					○			
5: Assuring impartiality and objectivity	✸	○	○	○	○		○			
6: Assuring transparency			✸				○			
7: Assuring statistical confidentiality and data security						✸				
8: Assuring commitment to quality		✸								
9: Assuring adequacy of resources	○									
Level C: Managing statistical processes										
10: Assuring methodological soundness		✸			○				○	○
11: Assuring cost-effectiveness					✸				○	
12: Assuring appropriate statistical procedures		✸			○					
13: Managing the respondent burden					✸					
Level D: Managing statistical outputs										
14: Assuring relevance	✸		○		○					
15: Assuring accuracy and reliability	✸				○					
16: Assuring timeliness and punctuality	✸				○					
17: Assuring accessibility and clarity	✸		○							
18: Assuring coherence and comparability	✸		○						○	
19: Managing metadata			✸						○	

UN National Quality Assurance Framework

1. Background

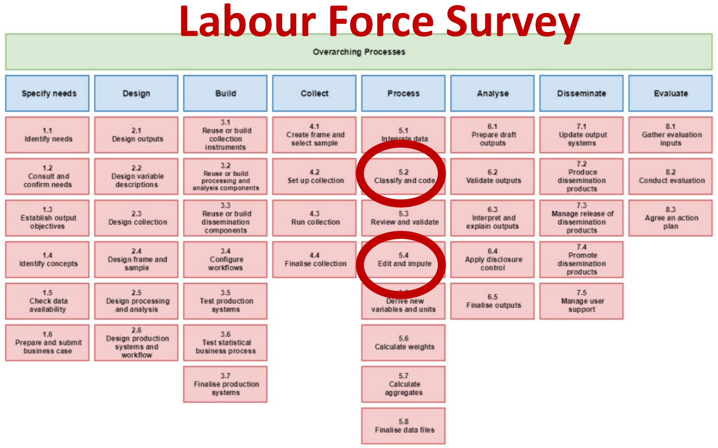
- Developed by UNECE HLG-MOS Machine Learning Project Work Package 2 – Quality Team in 2020
- Contributors: Siu-Ming Tam (Australia), Bart Buelens (Belgium-VITO) Wesley Yung (Canada), Gabriele Ascari and Fabiana Rocci (Italy), Florian Dumpert (Germany), Joep Burger (Netherlands), Hugh Chipman (Acadia University), InKyung Choi (UNECE)
- Final version in the Statistical Journal of the IAOS (2022)



2. Quality dimensions in QF4SA

- Why “Statistical” Algorithm? -> Applicable to both traditional statistical methods as well as ML methods
- Targeted for intermediate outputs, not necessarily for the final statistical output

Labour Force Survey



Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

Final statistical output (e.g., employment rate for 2021 Q4)

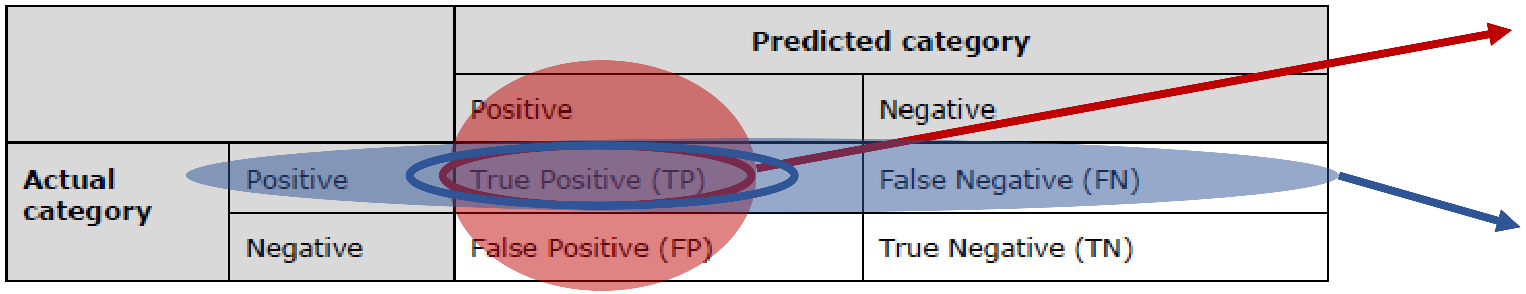
2.1. Quality dimensions - Accuracy

- Closeness of computations or estimates to the true values that were intended to measure
- Accuracy metric changes according to the process and to the target, when the focus is on unit wise predictive accuracy (often in ML application)

ID	Job description	Actual code	Predicted code	Result
1	I manage crane	8221	8221	Correct
2	Fork-lift	8222	8229	Incorrect
3	I drive lift trucks	8222	8222	Correct
…	…	…	…	….
3500	Plowing machine driver	8223	4133	Incorrect

2.1. Quality dimensions - Accuracy

- Closeness of computations or estimates to the true values that were intended to measure
- Accuracy metric changes according to the process and to the target, when the focus is on unit wise predictive accuracy (often in ML application)
  - For regression: RMSE (absolute or relative), etc.
  - For classification: accuracy, recall, precision and F1 score



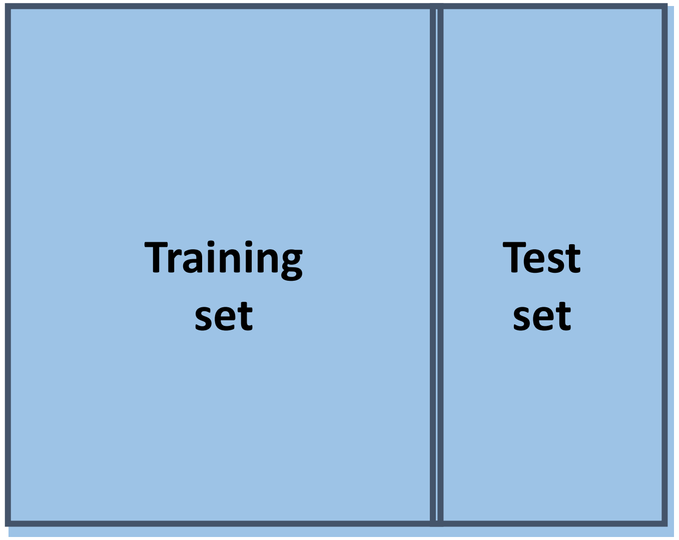
		Predicted category	
		Positive	Negative
Actual category	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Precision = True positives among all predicted positives

Recall = True positives among all actual positives

2.1. Quality dimensions - Accuracy

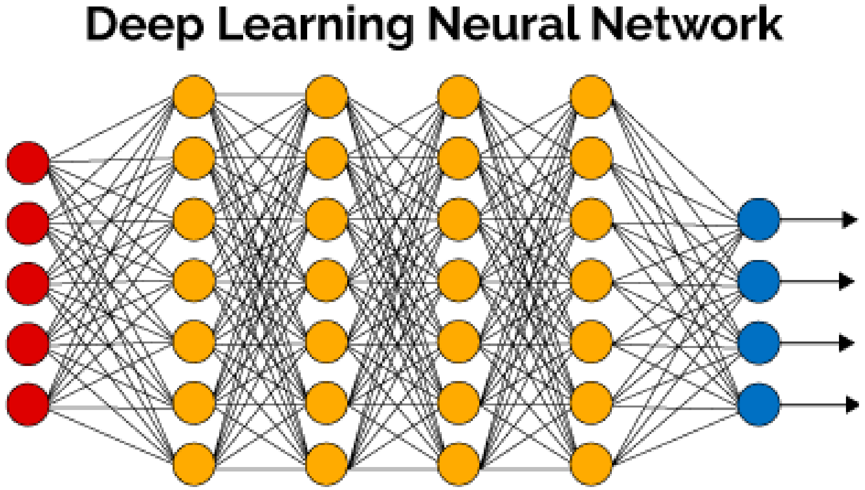
- ML methods often does not have as much restriction as traditional statistical methods
- Risk of overfitting to observed data
- Cross-validation scheme (split data set into training set vs. test set)



For realistic estimation of accuracy

2.2. Quality dimensions - Explainability

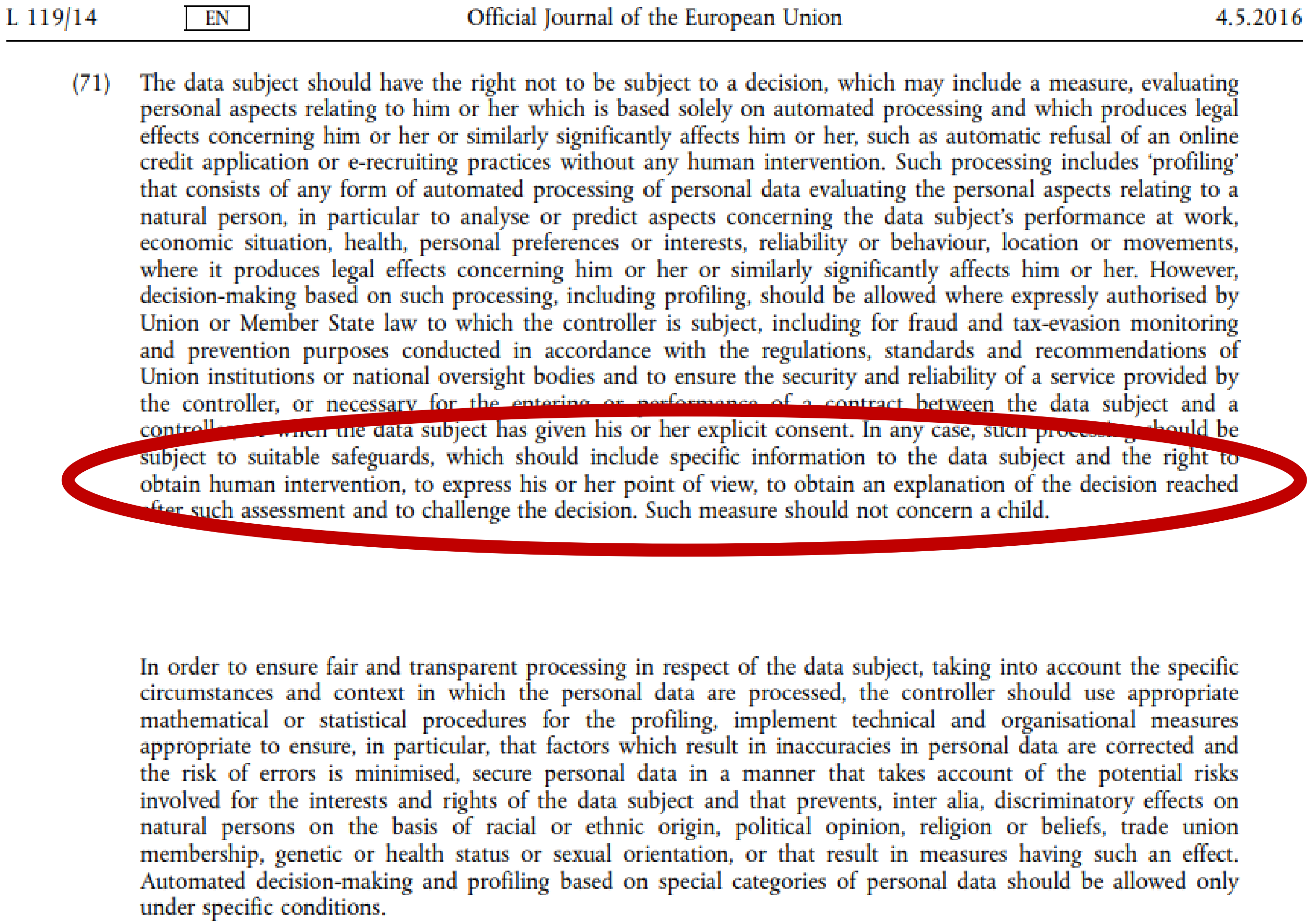
- Degree to which a human can understand how a prediction is made from a statistical or an ML algorithm using its input features
- Increased model complexity might improve accuracy but at the expense of model explainability



2.2. Quality dimensions - Explainability

E.U. General Data Protection Regulation (GDPR)

“…such processing should be subject to safeguards, which should include… the right to obtain an explanation of the decision…”

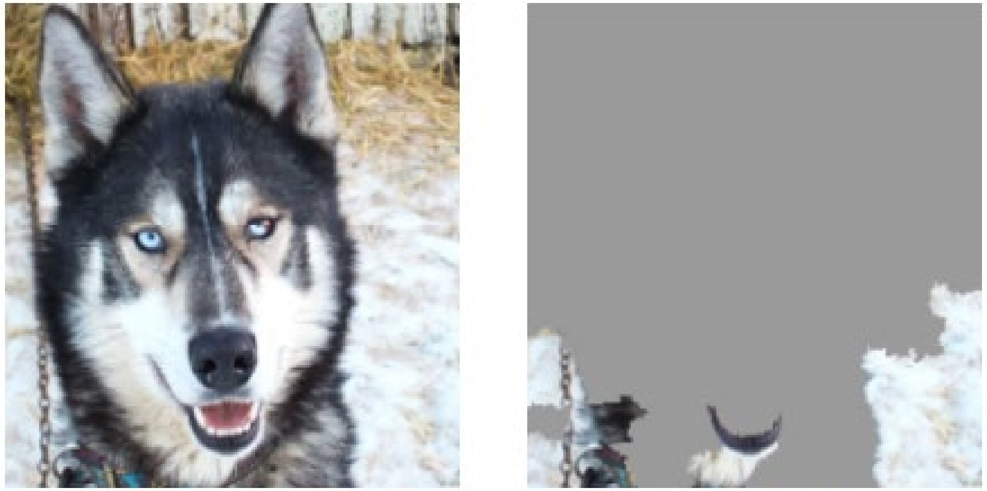


L 119/14 EN Official Journal of the European Union 4.5.2016

(71) The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes ‘profiling’ that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject’s performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measure should not concern a child.

In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.

2.2. Quality dimensions - Explainability



(a) Husky classified as wolf

(b) Explanation

Ribeiro et. Al. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier

2.3. Quality dimensions - Reproducibility

- Three types of reproducibility: methods reproducibility, inferential reproducibility and results reproducibility
  - Methods reproducibility is defined as the **ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results**
- Machine learning methods are often complex with a lot of parameters, hyperparameters, on top of dependency issues

2.3. Quality dimensions - Reproducibility

- How to make “reproducible”?
- Providing enough details about algorithms, assumptions and data so the same procedures could be exactly repeated, in theory or in practice, sharing analytical data sets (original raw or processed data), relevant metadata, analytical code and related software
- Analyses be repeated in-house and by another individual, who should be at arm’s length from the original researcher, to assess reproducibility

2.4. Quality dimensions - Timeliness

- The length of time between the reference period and the availability of information
- Also, recommend to consider
  - the length of time it takes to develop or put in place a process
  - the length of time it takes to process data
- The former can take long. But once in use, ML can process vast amounts of data in a short time
- Aspects to consider: Data cleaning, Preparation of training data, Evaluation of data quality, Model re-training

2. Quality dimensions – Cost effectiveness

- Degree to which results are effective in relation to the costs of obtaining them (e.g., RMSE reduction per unit cost)
- Fixed cost and on-going cost. Decomposing it into different cost components is useful to better assess potential savings and accuracy improvements against future ongoing costs. This would also help estimate the time needed to recoup the initial investment
- Some ML methods may introduce more cost than others

Cost component	**Type**	**Purpose**
IT infrastructure	Fixed	Necessary hardware and software
Initial staff training	Fixed	Training current staff; hiring new staff
Cloud storage	On-going	Cloud storage space
Quality assurance	On-going	Conducting quality assurance and control
…		

3. Summary

Different importance for different stakeholders at different stages

Quality dimension	**Method 1**	**Method 2**	**Method 3 (legacy)**
Accuracy	80%	85%	78%
Explainability	High (easy)	Low (hard)	High (easy)
Reproducibility	High (easy)	Middle	Low (hard)
Timeliness	High	Middle	Low
Cost effectiveness	High	Middle	Low

3. Summary

- Evaluating quality for statistical algorithms is multi-dimensional
- The proposed QF4SA presents five dimensions to help guide official statisticians when comparing different methods (ML and non-ML)
- The QF4SA is not a replacement for existing quality frameworks but is a supplement to them

Thank you for your attention