A Quality Framework for Statistical Algorithms

This paper was developed by the 'Quality Aspects' working group of the HLG-MOS Machine Learning project and consisted of: Wesley Yung – Chair (Canada), Siu-Ming Tam (Australia), Bart Buelens (Belgium), Florian Dumpert (Germany), Gabriele Ascari, Fabiana Rocci (Italy), Joep Burger (Netherlands), Hugh Chipman (Acadia University) and InKyung Choi (United Nations Economic Commission for Europe)

Executive summary

As national statistical offices (NSOs) modernize, interest in integrating machine learning (ML) into official statisticians' toolbox is growing. In the 2018 Blue Skies Thinking Network paper (UNECE 2018), two issues were identified: the potential loss of transparency from the use of "black-boxes" and the need to develop a quality framework. The goal of Work Package 2 of the High-Level Group for the Modernisation of Official Statistics' ML project was to address these two challenges. Many quality frameworks exist; however, they were conceived with traditional methods in mind, and they tend to target statistical outputs. Currently, ML methods are being looked at for use in processes producing intermediate outputs, which lead to a final statistical output. These processes include, for example, coding, imputation and editing, which do not directly produce statistical outputs, but whose results are used downstream to ultimately produce final outputs.

The working group is proposing the Quality Framework for Statistical Algorithms (QF4SA) to address the two issues identified in the Blue Skies Thinking Network paper. For the quality framework, the term "algorithm" is defined as a process or set of rules to be followed in calculations, derived from an assumed model and a predetermined set of optimization rules, for estimation or prediction. With this definition, the working group feels that both traditional and modern methods are covered. The proposed QF4SA comprises five dimensions:

1) explainability
2) accuracy
3) reproducibility
4) timeliness
5) cost effectiveness.

The QF4SA does not replace existing quality frameworks; it complements them. As the QF4SA targets intermediate outputs and not necessarily the final statistical output, it should be used in conjunction with existing quality frameworks to ensure that high-quality outputs are produced. The working group proposes the following recommendations:

1) It is recommended that all five dimensions of the QF4SA be considered when deciding on the choice of an algorithm, particularly when choosing between traditional and ML algorithms.
2) It is recommended that NSOs explore and use the methods outlined in the Explainability section to help users understand the relationship between input and output variables.
3) It is recommended that NSOs calculate the expected prediction error, as well as the prediction error, to protect against potentially poor-quality training data. In addition, it is recommended that high-quality training data be created when applying supervised ML algorithms.
4) It is recommended that NSOs use the highest-quality training data possible when applying prediction algorithms. This will also facilitate the estimation of expected prediction errors.
5) It is recommended that, as a minimum, NSOs take action to implement methods reproducibility. Inferential reproducibility should be carried out as well, when possible and desirable.
6) It is recommended that development and processing time be added to the commonly used concept of timeliness.
7) It is recommended that NSOs consider two aspects in particular when considering cost effectiveness: cheaper operating costs and time to recoup fixed costs.

1.0 Introduction

The aim of national statistical offices (NSOs) is to develop, produce and disseminate high-quality official statistics that can be considered a reliable portrayal of reality. In this context, quality is the degree to which a statistic’s set of inherent characteristics fulfills certain requirements (Eurostat 2014). These requirements are typically set out in a quality framework, which is a set of procedures and processes that support quality assurance within an organization and is meant to cover the statistical outputs, the processes by which they are produced, and the organizational environment within which the processes are conducted. Many widely accepted quality frameworks related to official statistics exist; for example, see the Australian Bureau of Statistics’ Data Quality Framework (ABS 2009), the United Nations’ National Quality Assurance Framework (UN 2019), Eurostat’s European Statistics Code of Practice (Eurostat 2017) and Statistics Canada’s Quality Assurance Framework (Statistics Canada 2017).

Modern methods such as machine learning (ML) are gaining popularity as tools for official statisticians. In combination with modern hardware and software, these methods allow official statisticians to process new data sources such as text and images, automate existing statistical processes, and potentially make inferences without a sampling design. With this increased interest, a quality framework must be considered for statistical processes where these methods could be used.

In a traditional estimation context, statisticians typically attempt to learn as much as possible about a scientific truth from observed data. As described by Efron (2020), the scientific truth can be represented as a surface, and the observed data can be thought of as observations on the surface obscured with noise. Efron calls this the **surface plus noise** formulation. For example, a simple linear regression uses a formulation $y = \beta_0 + \beta_1 x + \epsilon$, where the surface, or, in this case, the line, is represented as a linear function of a variable *x*, and the response value, *y*, is observed with noise ε. Based on a set of observations (or data), the parameters of the line are estimated (e.g., using maximum likelihood or ordinary least squares methods) to obtain the estimated surface.

ML, on the other hand, can be differentiated from the traditional estimation context by its focus on prediction as opposed to estimation. ML algorithms "go directly for high predictive accuracy and [do] not worry about the surface plus noise models" (Efron 2020). Rather than searching for a hidden truth about the underlying phenomenon that generated the data or characteristics of the population, ML primarily aims to make predictions about individual cases. Note that this does not mean traditional statistical algorithms cannot be used for prediction. Once the parameters of a regression surface, or line, are estimated (i.e.,$\hat{\beta}_0, \hat{\beta}_1$), they can be used to make a prediction for any given new data point, x (i.e.,$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$). For this reason, some traditional statistical algorithms are commonly found in the ML toolbox, to be used for prediction rather than estimation.

With different purposes, it is not surprising that traditional statistical and ML algorithms have different areas of application, where one performs better than the other. For example, city planners who are interested in understanding what factors cause congestion in certain districts may employ statistical methods that have a long history of successfully solving such problems. But companies providing real-time traffic services for commuters are more interested in predicting whether a certain route that a commuter is taking will be congested or not, and this is the area of prediction in which ML specializes. In situations where accurate predictions at the individual level are infeasible, ML methods may also see limited applicability. However, statistical methods can still deliver insight. For example, a statistical model such as a logistic regression allows the assignment of significance to individual predictors when modelling the occurrence of a disease, even if such an ML or classical statistical model cannot accurately predict which individuals will get the disease.

The popularity of ML in social media services, online shopping recommendations and search engine refinement is due to its ability to make predictions for individual cases. In the official statistics field, ML is becoming increasingly popular in areas where such individual prediction tasks are needed. These can be areas where these tasks used to be solved by traditional statistical algorithms (e.g., predicting whether a certain record needs editing) or by manual work (e.g., predicting to which category an open-ended response or satellite imagery pixel should be classified). This popularity may be because machine learners accept more complex models than traditional statisticians, and this can lead to higher predictive accuracy.

ML is a relatively new tool in the official statistics field. While there is a growing body of work on the methodological aspects of ML, less has been done on the quality aspects. Commonly used and accepted quality concepts may require re-evaluation through ML perspectives. For example, the United Nations' National Quality Assurance Framework states, "the accuracy of statistical information reflects the degree to which the information correctly describes the phenomena it was designed to measure, namely, the degree of closeness of estimates to true values" (UN 2019). While this accuracy is often considered as how accurately statistical estimates describe characteristics of the underlying population (e.g., unemployment rate estimate based on the Labour Force Survey), accuracy for ML can also mean how accurate predictions are for individual cases in an intermediate processing task as part of the entire production process. Also, unlike manual classification done by humans, ML methods are scalable but may require initial development and investment. This affects cost effectiveness and timeliness of the end product in a different way than existing methods. The specificity of ML methods requires new quality dimensions (e.g., explainability and reproducibility) that are not considered in existing quality frameworks.

The goal of this document is to propose the Quality Framework for Statistical Algorithms (QF4SA) to provide guidance on the choice of algorithms (including traditional algorithms) for the production process. Throughout this document, we define an algorithm as a process or set of rules to be followed in calculations, derived from an assumed model and a predetermined set of optimization rules, for estimation or prediction. Statistical algorithms are those used within a statistical context. We purposely use the terminology **statistical algorithm** as it covers both traditional and modern methods typically used by official statisticians. It is impossible to talk about algorithms without thinking of data. However, throughout this document, we do not address data explicitly, but we do recognize that there is an important interplay between algorithms and data. In particular, all quality measures proposed are conditional on the data that are available.

Under the QF4SA, we propose five quality dimensions: explainability, accuracy, reproducibility, timeliness and cost effectiveness. Most of these dimensions are considered in existing quality frameworks for statistical outputs, but, in the QF4SA, they apply specifically to statistical algorithms that typically produce intermediate outputs. For example, classification and imputation are processes in the production chain whose results are used in subsequent steps. The QF4SA concentrates on these intermediate outputs, as ML algorithms seem to be used, for now, in these contexts. The QF4SA'sdimensions are defined below:

Explainability

Explainability is defined as the ability to understand the logic underpinning the algorithm used in prediction or analysis, as well as the resulting outputs. Explainability is greatly assisted by depicting the relationship between the input and output variables and providing the necessary information on the methodology underpinning the algorithm.

Accuracy

Slightly different definitions of accuracy are given in several internationally accepted frameworks. The definition proposed for the QF4SA can be summarized as follows: the accuracy of statistical information refers to the degree to which it correctly describes the phenomena it was designed to measure; i.e., it is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure.

Reproducibility

At the basic level, reproducibility is defined as the ability to replicate results using the same data and algorithm originally used. This is known as methods reproducibility. At a higher level, it is defined as the production of corroborating results from new studies using the same experimental methods (results reproducibility), or similar results using different study designs, experimental methods or analytical choices (inferential reproducibility).

Timeliness

For the QF4SA, timeliness is defined as the time involved in producing a result from conceptualization to algorithm building, processing and production. A distinction should be made between timeliness in development and production, with the former generally taking longer than the latter.

Cost effectiveness

Cost effectiveness is defined as the degree to which the results are effective in relation to their cost. It is a form of economic analysis that compares the relative merits of different algorithms. For this purpose, cost effectiveness can be defined as the accuracy (e.g., measured by the mean squared error [MSE] or F1 score) per unit cost. Note that the total cost of doing the work—including fixed costs, such as infrastructure and staff training, and ongoing costs, such as production costs—should be taken into account.

It could be argued that there are other more appropriate definitions for these dimensions, but the purpose of the proposed quality framework is to open a dialogue on what official statisticians should think about when comparing statistical algorithms, be they traditional or modern. In what follows, we elaborate on each of the dimensions and propose aspects of each to consider when comparing algorithms.

2.0 Explainability

2.1 Description of explainability

In the QF4SA, explainability is defined as the **degree to which a human can understand how a prediction is made from a statistical or an ML algorithm using its input features**. Throughout this document, we use the term "feature" to represent individual independent variables that are inputs. This is synonymous with "explanatory variable", "independent variable" or "regressor" in more traditional contexts. Note that explainability concerns the relationship between input features and the predicted output rather than the "mechanical" understanding of the algorithm. For example, "finding a hyperplane separating data points by class" is a mechanical understanding of a support vector machine (SVM), while an explanation such as "the higher the value of feature *X*, the more likely the output is classified as category *Y*" provides an understanding of how the input feature is related[1] to the output. Note that a prediction can be explainable but might not be interpretable even with domain knowledge, i.e., in the above example, lack of a scientific explanation for why *X* produces the output *Y*. An ML algorithm is explainable as long as subject-matter experts and other users can assess the logic of the way the algorithm makes a decision (see "Importance of explainability" below)—for example, a type of chemical whose effect on the output is not well known but that turns out to be an important factor. Explainability can therefore be pictured as a concept between mechanical understanding and interpretability.

Predicted values, or predictions, from statistical models are often considered more explainable than those from ML models because statistical models tend to be more explicit in linking inputs to outputs. For example, regression coefficients explain the direction and strength of the relationship between a feature and the output. However, this is not always the case. The explainability of a regression model becomes unwieldy in a generalized linear mixed model with many (potentially transformed) features, their interactions, their effect on regression coefficients and a non-identity link function. On the other hand, the explainability of a prediction from a deep decision tree (an ML algorithm) is straightforward.

[1] A relationship revealed in any model trained on observational data does not imply causation. For instance, increasing the value of feature *X* through a subsidy or tax benefit may not be a successful policy-making strategy to promote category *Y*.

While predictions from a single decision tree are explainable (e.g., the prediction of instance i is $\hat{y}_i = y_i$ because feature $X_{1i} > x_1$and feature $X_{2i} > x_2$), predictions from a random forest—combining predictions from hundreds of decision trees—are less explainable because a user cannot discern how input features lead to the output. Given enough data, more complex ML algorithms, such as (deep) neural networks, may outperform simpler algorithms in terms of prediction error because they can better learn nonlinear relationships and interactions. Combining multiple algorithms through bagging, boosting or stacking may further reduce prediction error and prevent overfitting. However, **improved algorithm performance through increased complexity comes at the expense of explainability** because as an algorithm becomes increasingly complex, it is often more difficult to explain.

2.2 Importance of explainability

Explainability is important to gain users' **trust** in ML algorithms, as they are often considered "black-boxes." Understanding how an ML algorithm makes decisions can increase users' trust since they can relate the behaviour of the ML algorithm to their prior knowledge and internal logic. We do note that explainability might be user-specific. For a statistical organization using ML algorithms, users can include statisticians who may not be familiar with ML methods and subject-matter experts in the organization; data providers in partner organizations; data users such as the general public, academics and policy makers; and data scientists developing ML algorithms.

Understanding how algorithms make certain predictions can shed light for users on hidden patterns within the data that humans cannot easily perceive. This could provide **new insights** about phenomena (for subject-matter experts) and help improve the performance of the algorithm itself (for ML developers).

While high prediction accuracy indicates that an ML algorithm performs well, an algorithm can make a correct decision for the wrong reasons. For instance, Szabo (2019) describes an example where an automatic system developed to predict a patient's risk of pneumonia based on X-ray images turned out to have simply learned the type of X-ray machine. The reason was that doctors usually took X-rays with portable X-ray machines for patients in critical condition and in urgent need of diagnosis, whereas patients without serious conditions were sent to a radiology department where their X-ray would be taken with a different type of X-ray machine. If an algorithm is a black-box, the outputs could, at best, be of limited use to the user and, at worst, be misunderstood in critical decision making. In some circumstances, this could have an impact on human life. Therefore, by requiring some human intervention, explainability can serve as a **safeguard** that machines are making correct decisions for the right reasons.

Explainability draws great attention to developing **fair**, **accountable**, **transparent** and **ethical** artificial intelligence. When decisions made by a machine have a direct and significant impact on the daily lives of people (e.g., medical diagnostics, autonomous driving, fraud detection, social credit), it is important to ensure that such decisions are made fairly and ethically. For example, if an ML model developed for credit approval with hundreds of features happens to make decisions based mostly on certain demographic variables, the algorithm is likely to be considered unethical. Therefore, it should be corrected before deployment, regardless of how accurate its prediction is. ML algorithms are often considered neutral and independent as they make decisions solely based on data and free of human bias. However, because of the very fact that they "learn" from data, accidental bias in data can be perpetuated by ML algorithms if careful checks and balances are not in place. Given the increasing awareness that human subjects should be provided with an "explanation of the decision reached [through automated processing]" (EU 2016), NSOs, as public agencies, should be aware of these issues with the use of ML. For example, could the output of an ML process using many features identify unique individuals in a population?

2.3 Making predictions explainable

Explainable ML, or explainable artificial intelligence, is a recent but very active field of research. A multitude of methods, each with its own benefits and caveats, have been proposed to make predictions from black-box algorithms more explainable. Note that these methods do not directly make the ML algorithms more explainable. Instead, they make predicted results more explainable, and this sheds lights on the algorithm's behaviour, thus improving understanding of how the algorithm works. The objective of this subsection is not to provide technical or methodological details of those methods but to introduce briefly a few existing methods developed in the ML community as a starting point. Readers who are interested in further information are encouraged to consult the resources listed in the references (e.g., Arrieta et al. 2020; Vilone and Longo 2020; Molnar 2019; and Bhatt et al. 2020).

An important group of explainability methods shows the **importance of features**, by visual plots, quantitative measures or surrogate models. One way to assess feature importance is to plot how the model prediction of an instance changes when the value of one feature is changed. For example, assume there are *p* features (X_1,...,X_p) and one output variable (Y). For each instance *i*, changing the value of X_{1i}, while fixing the value of all other features, will create a line of predicted values that shows how the individual prediction $\hat{y}_i$changes with the value of feature X_1. Combining all (or a sample of) instances together yields an **individual conditional expectation (ICE)** plot for feature X_1.

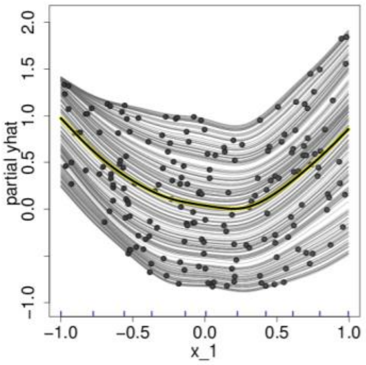

Figure 2.1 Example of individual conditional expectation plot

(Source:Goldstein et al. 2014.)

A **partial dependence plot (PDP)** averages over all instances to show the overall marginal effect of a feature on the model prediction. While ICE plots and PDPs are intuitive and easy to implement, they assume that the feature of interest, plotted on the X-axis, is uncorrelated with other features. This might not be true in real situations.

Another way to assess feature importance without retraining the model is to measure the increase in prediction error when a feature is permuted, i.e., its values are shuffled to break up the relationship between the feature and the outcome. A **surrogate model** is an explainable model that approximates the relationship between the features and the outcomes predicted by a black-box model. The surrogate model provides an explanation for the prediction by the black-box model. **Local interpretable model-agnostic explanations** are an implementation of a surrogate model for the purpose of explaining a single prediction. New instances and their black-box predictions are generated around the instance of interest. An interpretable model is trained on the generated data, weighted by their distance in feature space to the instance of interest. For example, the figure below shows a complex relationship between the two-dimensional feature space (X-axis and Y-axis) and binary output class (red and blue). An instance of interest is chosen (bold red cross), new instances are drawn from the feature space and their output values are predicted (crosses and points), and an explainable model (dashed line) is fit to the generated data, weighted by their distance from the instance of interest (size of crosses and points).

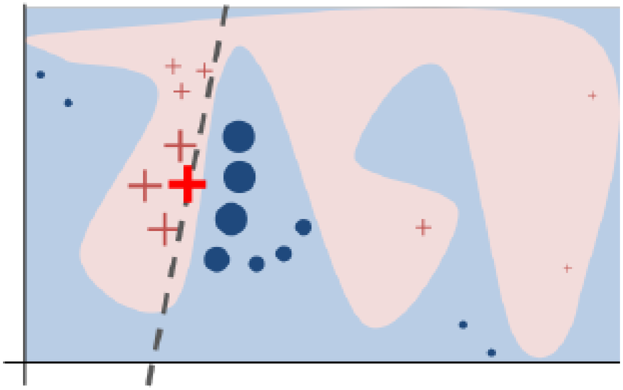

Figure 2.2 Example of local interpretable model-agnostic explanation

(Source: Ribeiro et al. 2016.)

The **Shapley value** is a measure of the contribution of a single feature value to the prediction of a single instance. It is calculated by comparing the predictions between different values of the feature, averaged over all (or a sample of) possible combinations of values for the other features. The contributions sum to the difference between the individual and average prediction.

Another group of explainability methods finds[2] data points in feature space that are intended to serve as the following:

- **Counterfactual example:** This is a data point that is as close as possible in feature space to the instance of interest but with a different predefined outcome. For example, assume that a description of a work-related injury is "I cut my finger while chopping something on a wood board," and the occupation of the person is classified as "a cook." However, if the description had been "I cut my finger while **carving** something on a wood board," the outcome would have been "a sculptor." The change in feature space between the predicted outcome and the counterfactual (e.g., "chopping" for "cook" vs. "carving" for "sculptor") is a counterfactual explanation.
- **Adversarial example:** This is an instance when one or more feature values have been slightly perturbed in a way that the right prediction turns into a wrong prediction (e.g., making an image classifier mislabel an image of a stop sign by adding a sticker to it). Although designed to mislead a trained image classifier, adversarial examples can be used to improve model security and robustness, and thus explainability.
- **Influential instance:** This is a data point in the training set that considerably affects the performance of the algorithm when deleted. For some algorithms, influence functions can approximate an instance's influence without the need to retrain the model.

[2] We focus on describing the data points of interest but omit how to find those data points through optimization of loss functions.

Traditional statistical algorithms employ intuitive formulations, which produce results that are innately explainable. ML algorithms may have higher predictive accuracy than these traditional methods, but, because of their complexity, they are often considered incomprehensible black-boxes. This can hamper the acceptance of ML in statistical organizations. Therefore, as ML becomes more common in the production of official statistics, the QF4SA recommends that if complex algorithms are used in any phase of output production, the official statisticians putting these algorithms in place must not only focus on minimizing the prediction error but also make a strong effort to achieve explainability by adopting some of the methods outlined above.

3.0 Accuracy

In the context of ML, we note that there may be some confusion when discussing accuracy: the term "accuracy" is used for a specific performance indicator in classification and ML (namely the fraction of correctly classified data points). However, in this section, we will present a much wider concept of accuracy and list several indicators to calculate it accordingly, with a special focus on ML.

We also note that, depending on the variables involved, measures of accuracy could take on different forms. For continuous variables, the MSE may be used to measure accuracy, with the bias squared component to quantify the effects of measurement errors. Other measures include mean absolute deviation, mean absolute relative deviation and distributional measures such as the Kullback–Leibler deviation. For categorical variables, accuracy may be measured by the misclassification rate or other measures of agreement between predicted and observed classes (e.g., informedness, markedness, F1 score, Matthews correlation coefficient or Cohen's kappa), a deviance (-2*log likelihood), or the area under the receiver operating characteristic curve.

3.1 Accuracy in official statistics

Accuracy has many attributes, and, in practical terms, there is no single aggregate or overall measure of it. Of necessity, these attributes are typically measured or described in terms of the error, or the potential significance of error, introduced through individual sources of error.

Accuracy can be said to relate to the concept of measuring the distance between the estimate (output) and the true value in an appropriate way. The closer the estimate is to its true value, the more accurate it is. We note that the deviation may be structural (bias) or random (variance).

For every framework, qualifying comments are common. For instance, the Australian framework states, “Any factors which could impact on the validity of the information for users should be described in quality statements” (ABS 2009). The Canadian framework states, “It should be assessed in terms of the major sources of errors that potentially cause inaccuracy. The accuracy of statistical estimates is usually quantified by the evaluation of different sources of error, where the magnitude of an error represents the degree of difference between the estimate and the true value” (Statistics Canada 2017).

These comments relate to the concept of measuring the distance between the estimate and the true value of the target parameter and refer to the closeness between the values provided and the (unknown) true values. This difference is called the error of the estimate, and “error” is thus a technical term to represent the degree of lack of accuracy.

3.2 Importance of accuracy

The mandate of many NSOs includes developing, producing and disseminating statistics that can be considered a reliable portrayal of reality. To ensure the high quality of these statistics, most NSOs have developed quality frameworks that cover the statistical outputs, the processes by which they are produced and the organizational environment. One of the most important components of every quality framework is accuracy, which is related to how well the data portray reality and has clear implications for how useful and meaningful the data will be for interpretation or further analysis. The concept of accuracy is defined across several frameworks in similar ways; the common fundamental notion is the closeness of the estimate to the true value.

Many measures of accuracy are available, each tailored to the particular estimation method being used and the situation (e.g., the type of data, the type of target parameter). Therefore, measures of accuracy can change according to the process and the target of the estimator. This target may refer directly to (G1) the data elements or (G2) aspects about the distribution, as in the case of imputation. In addition, a common objective of statistical surveys is to estimate a set of parameters of the target finite population. Therefore, within a quality framework, (G3) the accuracy of the estimates of these parameters is generally also considered a key measure of quality. In all of these cases, the purpose of the measure is to quantify the closeness of the estimate to the true value.

It is important to underline that the existing literature on the performance of an algorithm suggests considering two different aspects when evaluating an estimator (e.g., Hand 2012):

a) In choosing the estimator for the job, one must consider the choice of predictor variables, the estimation of parameters, the exploration of transformations and so on. In this view, when choosing among different estimators, a performance comparison is necessary to choose the most efficient one for the job.

b) After an estimator has been chosen, the estimator's ability to predict the true values of new data must be assessed.

In official statistics, it is necessary to add an additional aspect to point (b) above:

c) When the final estimate is released, an estimate of its uncertainty is required.

Therefore, the question naturally arises about which method should be adopted for a particular problem. The answer, of course, depends on what is important for the problem; different estimation methods have different properties, so a choice should be made by matching these to the objective.

3.3 Accuracy of supervised machine learning for classification and regression

As defined before, accuracy is meant to measure the closeness of an estimate to the true value. This means it depends on the estimation method under study. Therefore, before going into detail on measures of accuracy, we first set the context of how ML algorithms are typically used.

3.3.1 Training, validating and testing principle

To set the context, it is important to describe, in general terms, how the process of estimation and prediction is performed within a supervised ML approach. Suppose that there is a set, *S*, of labelled data $S: \{(x_1, y_1), \dots, (x_N, y_N)\}$, which belong to two spaces, i.e.,$x_i \in \boldsymbol{W}$and$y_i \in \boldsymbol{Q}$. That is, *S* is a set of observations of given variables *X* and *Y* that take on values over the given spaces. In ML, the existence of a function linking the variables in the two sets is presumed,

$$Y = f(X).$$

An ML algorithm estimates the mapping function f (with $\hat{f}$) from the input to the output. The goal is to approximate the mapping function so well that, when there are new input data (*X*), it is possible to predict the value of the output variable (*Y*) for these data. Depending on the nature of the spaces (and thus the variables within), we differentiate the task as follows. If the output space consists of a finite number of elements, then the task is called classification. Otherwise, the task is called regression. In less technical terms, if the output variable is quantitative or numeric, the learning task is called regression; if it is qualitative or categorical, the learning task is called classification.

Regardless of whether the task is regression or classification, ML algorithms will attempt to learn the relationship between *X* and *Y* based solely on the available data observed in *S*. As a result, ML algorithms can be much more flexible than traditional modelling methods as they do not tend to presuppose particular relationships between *X* and *Y*. Of course, as algorithms become more flexible, the problem of overfitting must always be kept in mind, i.e., the possibility that a learned model fits very well on the observed data (perhaps because it even interpolates the data) but generalizes poorly to as-yet-unseen data. Using pre-specified models with controls to avoid unnecessary complexity (e.g., very high dimensional polynomial terms)can reduce the danger of overfitting. However, ML simply tries to best estimate the mapping function, so it does not have such a restriction. Usually, the class of possible models is much larger and can contain high-order polynomials, which are susceptible to overfitting to the observed data. Regularization, stopping rules and the evaluation of the learned model on test datasets that have not been used during the learning process are schemes to deal with this potential problem, and they help improve the generalizability of estimators or predictors. Therefore, an ML model should be learned in the following way: the set of available data is split randomly into several (ideally independent) subsets, $\mathrm{S} \equiv \mathrm{A} \cup \mathrm{V} \cup \mathrm{T}$. For every set, the data in S are split accordingly, i.e., some (x_i, y_i) belong (not necessarily only) to A, some to V, some to T (see Figure 3.1). Note that Figure 3.1 is for illustrative purposes and does not suggest an optimal number of validation or testing sets or a ratio between the two.

- The first set, A, is for training the model (blue box).
- The second set (or sets), V(orange boxes), is used to find the best model parameters (e.g., the *k* in a *k*-nearest-neighbour approach, or the cost parameter [*C*]in an SVM approach).
- The third set (or sets), T (green boxes), is used to simulate what will happen when we apply the finally learned model to new, as-yet-unseen data.

The random attribution of units to *A*, *V* and *T* is important to avoid concept drift, as explained by Efron (2020). The final estimate $\hat{f}$ of the function f is always obtained on the training set$A$ (in combination with validation sets or not) and assessed on the test set or sets. Having more than one orange and more than one green subset not only allows for point estimates for the accuracy measures, but also enables an estimate of their variance.

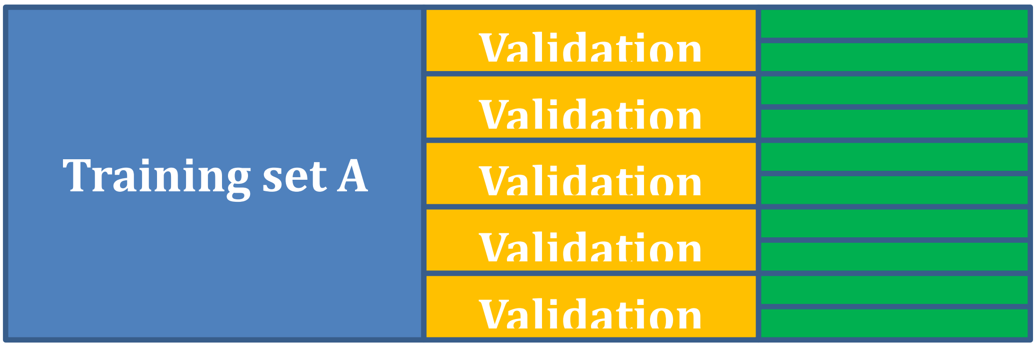

Figure 3.1 Training, validation and test sets

The set-up described in Figure 3.1 is the best way to split the set *S*, but, for various reasons, practitioners may choose other ways. Often, when there is just one validation set (in orange), bootstrapping or cross-validation is used on this single validation set to simulate the ideal situation, in which there are multiple validation sets. At times, because of a lack of data, there is no validation set. In this situation, if optimal values for the parameters have to be found, this can be done via cross-validation or bootstrapping within the training data.

The simplest and most commonly used version is to learn some models based on one training set (perhaps with cross-validation to specify some parameters) and to test them on only one test set (or bootstrap samples are created to provide multiple test sets to approximate the situation above). By repartitioning *S* into *A* and *T* multiple times, we have the opportunity to train and test the different algorithms or parameters of the algorithms on multiple datasets, thus showing us their performance in choosing the most efficient algorithm or parameters.

3.3.2 General approach for assessment of accuracy

Following Hastie et al. (2009), we will provide some details on this topic. Once the estimate of the function, $\hat{f}$, is obtained on the training sample *A*, a loss function is usually considered in order to calculate the error of predicting with regard to observed values. Typical examples of loss functions for when the variable *Y* is numeric include the following:

$$L\left(Y, \hat{f}(X)\right) = (Y - \hat{f}(X))^2 \ (squared\ error)$$

$$L\left(Y, \hat{f}(X)\right) = \left|Y - \hat{f}(X)\right| \ (\ absolute\ error).$$

In a classification situation, when the variable *Y* is discrete, a simple loss function is the zero-one loss function given by

$$L\left(Y, \hat{f}(X)\right) = I(Y \neq \hat{f}(X)).$$

The assessment of an ML approach goes through the evaluation of a loss function that indicates the ability of the given algorithm to perform as well as possible in predicting the output given new data, as follows:

$$Err_A = E\left[\, L\left(y, \hat{f}(x)\right) \middle|\, A\,\right]$$

where (x,*y*) is a data point drawn from the joint distribution of *(X,Y)*. Note that this error is conditional on the training set, *A*, and this error is often estimated by

$$\widehat{Err}_A = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \hat{f}(x_i)\right)$$

where (x_i, y_i) are points in the test set *T* of size *n*. If we average over all possible training sets, *A*, we obtain the expected error

$$Err = E_{A\subset S} E\left[\, L\left(Y, \hat{f}(X)\right)\middle|\, A\right].$$

The choice of which error to calculate depends on the situation at hand. If the general performance of an ML algorithm is of interest, it is necessary to estimate *Err*, which gives some protection from a poorly constructed training set. Furthermore, *Err* gives an impression of the robustness of an approach when the input data vary slightly. Fortunately, cross-validation seems to estimate *Err* well (Hastie et al. 2009) if *A* is a representative dataset.

However, when a particular ML model (a predictor, an estimator) has already been learned, it has been learned based on a concrete training dataset *A*, so Err_A has to be calculated to get an impression about the future performance of this particular ML algorithm. This is particularly useful when comparing algorithms but one should note that if the dataset A is not representative, then *Err* should also be calculated to get an idea of the performance of the ML model(s) across different possible training datasets.

3.3.3 Variance

One common point of criticism of ML concerns the question of how to measure the uncertainty of the outputs. Besides the closeness of computations or estimates to the exact or true values (which can, for example, be expressed by the bias), statisticians also consider the variance of an estimator. This can be used to calculate confidence intervals, or the uncertainty of predictions, which can be used to calculate prediction intervals. In parametric model-based statistics, formulae are usually available for these quantities. The estimated variances of some traditional estimators can be written down in closed formulae; for example, if logistic regression is used, confidence intervals for the parameters and prediction intervals for the predictions themselves are available. As there is currently a lack of mathematical statistical theory for some ML algorithms, results like these cannot be produced at this time for those approaches without making additional assumptions. We note that assumptions are also required in traditional methods. However, in the case of binary classification, Scholtus and van Delden (2020) have derived estimators of bias and variance for estimates of counts, proportions, differences of counts and growth rates. Their context assumes a binary classifier is used, and the resulting classification is used to produce the estimates mentioned above.

In the context of both ML and traditional statistics, resampling methods such as the jackknife (Quenouille 1956), cross-validation (Stone 1974) and the bootstrap (Efron 1979) have been developed and can be used to quantify the uncertainty on the three levels, (G1) to (G3), mentioned above. Wolter (2007) presents an introduction that focuses on the survey sampling context, while studies in the classification and regression context include those of Kim (2009) and Borra and Di Ciaccio (2010), respectively. Of course, the suitability of using these resampling methods for the algorithm and data at hand has to be demonstrated before they are used. This is emphasized here because there are situations where, for example, the empirical bootstrap does not deliver suitable results (e.g., Bickel and Freedman 1981). However, their examples of bootstrap failures are unlikely to occur in official statistics. Care, however, needs to be exercised to ensure that the dataset to which resampling methods are applied are representative of the population on which valid inference is to be made. Next, we present some details on how cross-validation and bootstrap samples can be used to evaluate statistical algorithms.

K-*fold cross-validation*

K-fold cross-validation uses part (or a fold) of the data to train the model and another fold to test (or validate) it. This is done by splitting the data randomly into *K* roughly equal folds. The model is trained on *K-1* of these folds and then tested (or validated) on the *k*-th fold (the one not used for training), and the prediction error is calculated. This is repeated for *k=1, …, K*, and the *K* prediction errors obtained are averaged. More formally, the process is as follows:

- Let $\hat{f}^{-k}(X)$ be the prediction of *Y* based on the model obtained when the *k*-th fold is omitted.
- The estimated conditional training error based on using the *k*-th fold as test data is then

$$\widehat{Err}_A^{-k} = \frac{1}{N_k} \sum_{(x_i,y_i) \epsilon F_k} L\left(y_i, \hat{f}^{-k}(x_i)\right)$$

  where F_k is the set of N_k units in the *k*-th fold.
- Repeat for *k=1, …, K.*
- The estimate of the expected error is then

$$\widehat{Err}_{CV} = \frac{1}{K} \sum_{k=1}^{K} \widehat{Err}_A^{-k}.$$

Typical values for *K* are, according to the literature, 5 and 10.

As mentioned by Hastie et al. (2009, 249), evaluating the variability of the cross-validation error estimates is important. This can be done by calculating

$$\widehat{Var}_{CV}(\widehat{Err}_{CV}) = \frac{1}{K-1} \sum_{k=1}^{K} \left(\widehat{Err}_A^{-k} - \widehat{Err}_{CV}\right)^2$$

as an estimate of the variance of the expected error rate, *Err*. However, note that there is no unbiased estimator for the variance of the cross-validation estimator (Bengio and Grandvalet 2004).

Note that there are also the predictions, $\hat{f}^{-k}(x_i)$, for all of the *x_i*, so it is possible to formally calculate an estimate of the variance of prediction, when Y is continuous,

$$\widehat{Var}_{CV}(pred) = \frac{1}{N-1}\sum_{k=1}^{K}\sum_{(x_i,y_i)\in F_k}\left(y_i - \hat{f}^{-k}(x_i)\right)^2$$

where $N = \sum_k N_k$ is the total size of the set *S*. If the folds are of equal size, that is, N_k=*N/K*, then under a squared error loss function, $\widehat{Var}_{CV}(pred)$ is equivalent to $\widehat{Err}_{CV}$ except for a factor of *N* versus (*N-1).* It is also possible to look at the residuals, $\left(y_i - \hat{f}^{-k}(x_i)\right)$, to come up with something similar to empirical 95% prediction intervals, but, again, there is the limitation that the probability distribution of the cross-validation estimator is not known exactly. Vanwinckelen and Blockeel (2014) provide a critical discussion of cross-validation.

Bootstrap

For the bootstrap, we draw a simple random sample of *N* units with replacement from the original training set,

$$A_b = \{(x_1^*, y_1^*), \dots, (x_N^*, y_N^*)\}.$$

Let T_b be the set of units that are not selected in the *b*-th bootstrap sample, and let $\hat{f}^b(x)$ be the model obtained from the *b*-th bootstrap training sample. Use T_b to test the model and calculate the estimate of the error as follows:

$$\widehat{Err}_{T_b}^{b} = \frac{1}{N_b}\sum_{(x_i,y_i)\epsilon T_b} L\left(y_i, \hat{f}^b(x_i)\right),$$

where N_b is the number of units in T_b. This is repeated *B* times (say 100 or more), and the estimated expected error is then

$$\widehat{Err}_{BS} = \frac{1}{B}\sum_{b=1}^{B}\widehat{Err}_{T_b}^{b}.$$

The following quantities can also be calculated:

$$\widehat{Var}_{BS}\left(\widehat{Err}_{BS}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{Err}_{T_b}^{b} - \widehat{Err}_{BS}\right)^2$$

$$\widehat{Var}_{BS}(pred) = \frac{1}{N^*-1}\sum_{b=1}^{B}\sum_{(x_i,y_i)\in T_b}\left(y_i - \hat{f}^{-k}(x_i)\right)^2$$

where $N^* = \sum_{b=1}^{B} N_b$. In addition, prediction intervals can be calculated. Note that there are several situations where this “standard version” of the bootstrap is not suitable for estimating the variance of a quantity—for example, in time series. Other versions have been developed over the years (see, e.g., Wolter [2007, 194] for some references). The proposed variance estimators borrow heavily from traditional statistical methods, and their theoretical properties must be explored. Therefore, we caution users on their use until their properties are fully understood.

Resampling techniques are widely used to estimate variances in several situations. Many implementations (e.g., in R or Python) already provide them as standard procedures. Nevertheless, these techniques, which are surely necessary when using ML, also have some disadvantages and pitfalls. It is important to be aware of them. For example, it is highly recommended to carefully check which version of the bootstrap, under which assumptions, is appropriate for the individual problem at hand in official statistics. Also, be careful with inference (i.e., with statistics beyond exploration and description) using “confidence intervals” or “statistical tests” based on cross-validation.

3.4 Common measures for evaluating statistical algorithms or their results in machine learning

3.4.1 When estimating the target parameter (G3)

A conceptual framework for accuracy is the total survey error (TSE), which describes, ideally, the accumulation of all errors that may arise in designing, collecting, processing and analyzing survey data (Platek and Särndal 2001; Biemer 2010; Groves and Lyberg 2010). Commonly, the error components for a statistical process are listed as follows.

- Sampling error: This is the part of the difference between a population value and an estimate thereof, derived from a random sample, that is due to only a subset of the population being enumerated.
- Non-sampling error: This is an error in survey estimates that cannot be attributed to sampling fluctuations. Examples of non-sampling error include coverage error, measurement error, nonresponse error, processing error and model assumption errors.

Thus, the TSE accumulates all errors that can arise in sample design, data collection, processing and analysis of survey data, and it comprises both sampling and non-sampling errors. Technically, a number of measures may be used to indicate accuracy through the definition of a proper loss function. To quantify the TSE for the estimate of a (usually continuous) population target parameter, the most common metric used is the MSE (which is the square root of the sum of squared bias and the variance).

3.4.2 When the focus is on distributional accuracy (G2)

Distributional accuracy is an important aspect to consider when using statistical algorithms to impute for missing values. In addition to the prediction of the true unknown missing value, relationships between the variables, or distributional accuracy, must be considered. At least in higher dimensions, distributional accuracy cannot be measured easily by only one number. However, in the univariate situation, well-known tests (such as the Kolmogorov–Smirnov test) can check whether two distributions are significantly different from each other. In the multivariate case, interactions of the variables have to be considered. It might be necessary to calculate correlations between the dimensions, but also to calculate extreme values, moments and quantiles separately per dimension and to recombine them in a specified sense. If all this occurs within an imputation step, the number of broken plausibility or edit rules for imputed values (and, if possible, the impact on the downstream task) and the accuracy (ideally also the variance) of the estimation of the target parameters are also important indicators. When distributional accuracy is measured, the Jensen–Shannon metric appears to be appropriate, as outlined by Prasath et al. (2019), because of its versatility for handling multivariate distributions with continuous and categorical variables.

3.4.2 When the focus is on unit wise predictive accuracy (G1)

In the pilot studies undertaken within the ML project of the United Nations Economic Commission for Europe’s High-Level Group for the Modernisation of Official Statistics and the literature (e.g., Japkowicz and Shah 2011; Pepe 2003; Stothard 2020; Hand 2012), the following measures are commonly used to assess the success of ML algorithms:

- in the case of regression, root MSE (absolute or relative), mean error, mean absolute or relative error, R^2 or the standard error of regression
- in the case of classification, predictive accuracy, recall, precision and F1 score per class or on macro levels, G measure, Matthews correlation coefficient, and awareness of the consequences of the different misclassifications (see section 2.2 on the Importance of explainability).

The references mentioned above contain many more measures and more discussion about them. A critical point in the case of classification, for instance, is how sensitive measures are to class imbalances (see, e.g., Luque et al. 2019) or whether they need a prespecified threshold in the decision function. In the latter case, areas under curves are used to assess classifiers—for example, the area under the receiver operating curve and the area under the precision recall curve (see Hand [2012] for more). Note that when these measures are estimated for a particular task to evaluate how well the learned predictor works, these numbers are valid only for tasks in the same context and based on new data from the same distribution (or the same data-generating process) as the training and test data used for learning and assessing the predictor. This underlines the importance of having training and test data that are representative of the underlying population. This implies that the accuracy of an ML model must be continuously monitored and underlines the importance of having representative training and test data of the population being imputed for.

4.0 Reproducibility

4.1 Dimensions of reproducibility

According to a subcommittee of the U.S. National Science Foundation (Stodden et al. 2018) on replicability in science, "reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results…. Reproducibility is a minimum necessary condition for a finding to be believable and informative."

It is important to recognize the three dimensions of reproducibility, namely, methods reproducibility, results reproducibility and inferential reproducibility (Goodman et al. 2016).

- **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. This is the same as the minimum necessary condition described in the U.S. National Science Foundation subcommittee recommendation.
- **Results reproducibility** is defined as the production of corroborating results in an independent study (i.e., with new data) that followed the same experimental methods. This has previously been described as replicability.
- **Inferential reproducibility** is defined as making knowledge claims of similar strength from a study replication or reanalysis. This is not identical to results reproducibility, because not all investigators will draw the same conclusions from the same results, or they might make different analytical choices that lead to different inferences from the same data.

For the QF4SA, recognizing that it is not feasible for official statisticians to undertake new data collection to corroborate initial findings, it is **not** proposed to adopt results reproducibility in official statistics.

Consistent with the Fundamental Principles of Official Statistics, methods reproducibility has been invariably embraced by NSOs, and its adoption in the QF4SA when using statistical algorithms to produce official statistics is expected to receive overwhelming support.

For inferential reproducibility, as multiple algorithms can generally be brought to bear on data analysis, there would be multiple ways to reanalyze the data. Official statisticians, when deciding to use a particular algorithm with a decided set of assumptions for analysis, have to be reasonably satisfied that the results from the chosen analysis can be corroborated by analyses using alternative but applicable algorithms and assumptions. This is particularly important for analytical inferences where general assumptions inherent in the algorithms have to be made about the data.

What is the distinction between accuracy and reproducibility? Accuracy is about having large accuracy metrics (e.g., small MSEs for continuous variables or large F1 scores for categorical variables), given a dataset, associated with the algorithm. Inferential reproducibility occurs when the MSE or F1 score of the difference between results obtained from the same dataset—from different choices of study designs, experiments or analytical techniques—is not statistically significant. In other words, inferential reproducibility is an attribute to show whether we can get essentially the same result (within a margin of error, and using algorithms correctly), not whether that result is good.

The following example illustrates the difference between accuracy and reproducibility:

> Suppose response Y depends on predictor X1 but not X2. We observe Y and X2 and build a model to predict Y from X2. That model concludes (correctly) that X2 is irrelevant in predicting Y. In that case, we would have poor prediction accuracy (high MSE for predictions of Y). As the same inaccurate predictions of Y would be obtained using the same model and assumptions, the analysis is methods reproducible. In addition, the result that X2 is irrelevant in predicting Y is inferential reproducible because different models used to model Y on X2 will show the same result, provided these models are thoughtfully and correctly applied. For example, a bad choice of hyperparameter or other inappropriate modelling decision could lead a model to overfit and incorrectly infer a relation between Y and X2.

In the above example, we have shown that reproducibility is an attribute to show whether we can get the same result (methods reproducibility) or corroborating result (inferential reproducibility), not whether that result is good.

4.2 Importance of reproducibility for official statistics

Reproducibility builds and enhances trust in official statistics. The third principle of the Fundamental Principles of Official Statistics, accountability and transparency, adopted by the United Nations Statistical Commission in 2014, stipulates that “To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics." (UN 2014).

Gleser articulated an underlying rationale for reproducibility in 1996. When commenting on DiCiccio and Efron's (1996) seminal paper on bootstrap confidence intervals published in *Statistical Science*, Gleser said the "first law of applied statistics" is that "two individuals using the same statistical method on the same data should arrive at the same conclusion." (Gleser 1996).

In the academic world, to ensure this first law of applied statistics is followed, many journals have revised author guidelines to include data and code availability. For example, starting February 11, 2011, the journal *Science* requires the following:

> "All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*. All computer codes involved in the creation or analysis of data must also be available to any reader of *Science*. After publication, all reasonable requests for data and materials must be fulfilled. Any restrictions on the availability of data, codes, or materials, including fees and original data obtained from other sources (Materials Transfer Agreements),must be disclosed to the editors upon submission."

Trust is the currency of official statistics. While many factors contribute to building trust, an important one, as outlined in the first of the Fundamental Principles of Official Statistics, is impartiality. Impartiality can largely be demonstrated by transparency in the sources, methods and procedures used to compile official statistics. Such transparency allows independent analysts or researchers to assess the integrity of—and, where possible, reproduce and verify—published official statistics.

4.3 Demonstrating reproducibility

Those who develop statistical algorithms to compile official statistics (e.g., methodologists, data scientists and analysts) are encouraged to assess the methods and inferential reproducibility of their algorithms before adoption. Once the reproducibility dimension of the algorithms has been confirmed during the development stage, they can be put into production, and no further reassessment would be considered necessary.

Methods reproducibility refers to providing enough details about algorithms, assumptions and data so the same procedures could be exactly repeated, in theory or in practice. Documenting methods reproducibility therefore requires, at minimum, sharing analytical datasets (original raw or processed data), relevant metadata, analytical code and related software. For confidentiality reasons, NSOs generally cannot share identifiable raw data for independent analysis. It is therefore proposed that the analyses be replicated in-house and by another individual, who should be at arm’s length from the original researcher, to assess reproducibility.

For inferential reproducibility, methodologists should test corroboration of the results from the chosen algorithm with a small set of applicable algorithms and different assumptions. While there are no hard and fast rules to determine what constitutes corroboration, judgment should be applied when examining the results that are “different” from those of the chosen algorithm and assumption. For example, are the differences statistically significant, i.e., not due to random fluctuations? If they are, can they be explained (e.g., because of an improvement in efficiency), and can the explanation be supported by statistical theory?

Lastly, it is also proposed that the outcomes of methods and inferential reproducibility be documented for longevity and, where possible, published as part of the quality declaration statement normally released together with the official statistical output.

Clearly, reproducibility of statistical algorithms is fundamental for upholding trust in official statistical outputs. While three types of reproducibility are recognized in this section, we propose NSOs adopt methods and inferential reproducibility to support their choice of statistical algorithms in producing outputs.

5.0 Timeliness

5.1 Timeliness for statistical algorithms

The quality guidelines and frameworks of many NSOs (Statistics Canada, the Australian Bureau of Statistics, the Office for National Statistics, and the Organisation for Economic Co-operation and Development) define timeliness as the length of time between the reference period and the availability of information. The QF4SA is advocating for development and processing time to be considered as well as the normal timeliness measures. More broadly, the concept of timeliness should be expanded to cover the period of time between a need for data and the release of the information to meet that need. With the increased use of large datasets, the speed at which ML algorithms can be trained and run could lead to significant improvement in timeliness. This is particularly true for processes that are typically done manually, such as coding. Coding applications can be developed quickly using ML, particularly if past manually coded data can be used as training data. In addition to being fairly quick to set up, once developed, ML algorithms can process vast amounts of data in a short time. Compared with manual processes, ML algorithms could lead to significant savings in processing time.

5.2 Importance of timeliness

Official statistics are useful only when they are relevant, and this means that they need to be available in a timely fashion. Indicators of an economic downturn are not relevant if they are available only six months after the downturn has occurred. Many quality frameworks define timeliness as the length of time between the reference period and the availability of information. However, for the QF4SA, we consider two additional dimensions of timeliness:

- the length of time it takes to develop or put in place a process
- the amount of time it takes to process data.

These two dimensions deserve consideration, as we feel that ML can offer some advantages over commonly used methods that can lead to improvements in the commonly used definition of timeliness.

5.3 Responsibility of ensuring timeliness

Typically, statistical algorithms, be they ML or other commonly used methods, would be developed or chosen collectively by methodologists, data scientists and analysts. Depending on the problem at hand, the work should also involve informatics specialists and subject-matter experts. Timeliness considerations are most likely evaluated during the development phase of a project. However, during redesigns or in continuous improvement contexts, this aspect of timeliness should be considered.

5.4 Aspects to consider

Clearly, measuring the time required in production to develop, set in place and use something is straightforward. In this section, we list some aspects that need to be considered during the evaluation.

- Data cleansing

  It is highly likely that all potential methods will require that similar data cleansing be performed. However, if certain methods require specialized preparation of input data, for some reason, then this should be recorded.

- Informatics infrastructure

  If the method requires an informatics infrastructure that is not currently available, then the time required to set up such an environment should be considered. The time required to put it in place should not be underestimated.

- Preparation of training data

  Supervised ML algorithms require high-quality training data and, depending on the method, a large quantity of data can be required. Existing data should be considered for training data, if appropriate.

Note that some traditional approaches also need auxiliary data, which can be time-consuming to obtain. For instance, non-ML coding algorithms typically need a data dictionary that is complete, accurate and up-to-date, but that can be very time-consuming to create.

- Evaluation of data quality

  Many well-established methods have processes for evaluating data quality. For instance, a well-developed theory exists for variance because of imputation. However, new approaches may not have well-defined processes to estimate quality indicators and may rely on resampling-type algorithms (e.g., cross-validation or bootstrap) to evaluate quality. Depending on the algorithm, these resampling methods could take significant time to compute.

- Scalability of the approach

  As data sources continue to grow in size, the time required to process large datasets should be considered. Manual processes are not a viable choice when the number of records to process becomes large, so ML algorithms may be preferable.

6.0 Cost effectiveness

6.1 Cost effectiveness for statistical algorithms

Cost effectiveness can be defined as the degree to which results are effective in relation to the costs of obtaining them. Results in statistics are mainly measured in terms of accuracy; therefore, it is natural to link cost effectiveness to the accuracy dimension and try to measure it from this perspective. In this section, we will define cost effectiveness as the accuracy (measured by the MSE for continuous data and F1 score or similar metrics for categorical data) per unit cost.

This is an operative definition that makes comparisons between different methodologies possible. In the case of ML, an organization may compare the accuracy of an ML algorithm with the accuracy of a traditional method for the same statistical process, expressing both approaches in terms of their unit costs. The assessment of accuracy in ML is usually based on the consideration of a loss function; in traditional methods, uncertainty is expressed by the variance of an estimator, but resampling methods may be used as well (see Section 3.3).The same comparison could be made, of course, between two or more ML algorithms if the objective were to choose the most cost-effective one, all other aspects considered. However, some practical issues may need to be considered with this method, especially related to which costs should be included in the analysis.

Whenever a new method is introduced in a production process, an organization will have to face some initial expenses to implement it. Such costs may be broadly defined as fixed costs, as they usually represent costs that must be paid to launch the infrastructure for the new method. ML, which can rely heavily on the underlying information technology (IT) infrastructure, may pose some challenges in this regard. In fact, fixed costs for ML mainly include the IT-related costs for acquiring new software and hardware and the costs of training the organization's staff. These are different from the other category of costs that can be identified—ongoing costs—which derive from regular efforts to keep the whole system running and up to date. The following table lists the possible costs of an ML project. It may be useful to note that traditional methods also present fixed costs. However, NSOs have been investing in these over many years, so additional fixed expenditures are not usually required for them.

Table 6.1 Potential additional fixed and ongoing costs for machine learning adoption

Cost component	Type	Purpose
Information technology (IT) infrastructure	Fixed	Acquiring necessary hardware and software
Cloud storage	Ongoing	Acquiring necessary cloud storage space
IT maintenance	Ongoing	Maintaining IT infrastructure
Initial staff training	Fixed	Training current staff on ML; may include hiring new staff
Ongoing staff training	Ongoing	Keeping staff up to date withnew ML developments
Data acquisition	Fixed/ongoing	Acquiring and processing new data sources
Quality assurance	Ongoing	Conducting quality assurance and control

The details of these components will be explained later in this section. For now, it should be noted that ML methods, by themselves, are not necessarily more expensive than traditional methods. In some cases, as they generally rely on less theoretical assumptions than classical statistics, they could be even simpler to implement and could be applied to traditional datasets without much difficulty. In such cases, where big data are not included, ML methods may present few additional costs. The elements shown in the table can be considered a starting point for comparing ML and traditional methods; such comparison can be made by (a) analyzing whether the running costs for ML methods are cheaper than those of traditional methods or (b) computing the number of years to recoup the investment needed for the extra elements outlined in the table. Finally, we note that fixed costs should not be associated with a single instance of implementing an ML algorithm (unless the NSO intends to implement a single instance of ML). Fixed costs should be spread over the number of ML algorithms under consideration and future possible applications. At some point, the costs of new ML instances will be nil.

6.2 Advantages of cost effectiveness

The last decade has seen an explosion in data production, because of improvements in computer processing speed and innovations in communication networks. Official statistics have therefore been forced to compete with an increasing pool of data producers, while often being limited by tight budget constraints. Statistical offices are facing a challenge in meeting the required high-quality standards of official statistics with the resources that are made available to them. Cost effectiveness has guided many statistical institutes in recent years: the European Statistics Code of Practice, for example, dedicates its principle 10 to cost effectiveness, stating that resources should be used effectively. Current statistical processes may be revised to achieve the same or better levels of accuracy using sources or methodologies that would allow the organizations to save some costs; new data sources may be explored to save costs in data collection procedures. Indeed, cost effectiveness is one of the reasons behind the shift by NSOs from survey-centred data production to processes involving administrative and alternative sources of data. The introduction of ML can be seen as a further step in this evolution.

6.3 Organizational considerations

ML in official statistics is still a field under investigation, although it has shown promising results. However, every organization is different in terms of available budget and statistical production, so the convenience of introducing ML into current production has to be looked at on a case-by-case basis. If an organization is new to ML algorithms and to big data sources in general, it would probably need to implement a suitable infrastructure from the start. Therefore, it will have to take into account the start-up costs and evaluate them against its budget, the cost of the current production, and the expected accuracy and timeliness improvements. Fixed costs may represent the main challenge in this case and may take a toll on the organization's budget, but they also have to be compared with the future savings that ML would grant. As a result, fixed costs could actually be considered an investment that would allow greater savings in the future. Such savings may depend on the characteristics of the statistical production itself, as some processes may be more suitable for a migration toward ML than others. A given organization may be involved in many projects that can easily—and beneficially—adopt an ML approach, while another organization may have too few such projects, in which case the initial investment would be harder to justify.

6.4 The potential costs of machine learning

As can be seen in Table 6.1, IT-related and staff-related costs are a big part of the costs linked to the adoption of ML. To illustrate these, two of the main advantages of ML methods—scalability and automation—are introduced.

Scalability implies that a procedure can be applied with no or few modifications to a larger scale—for example, to a bigger data source with a greater set of units or features. As noted earlier, ML methods per se do not necessarily require any additional effort in terms of computation or resources. However, when used in conjunction with big data, they can quickly become computationally intensive. ML algorithms are often based on iterative methods and, of course, the better the hardware, the faster such iterations will be. An organization's existing infrastructure may require some adjustments (e.g., central processing units, graphics processing units, storage space) before it can be used for computationally intensive operations or large datasets. Furthermore, IT costs should also include the resources needed for cloud storage and computation in the cloud, which are usually ongoing costs. In conclusion, when planning to introduce ML in a statistical process, an organization could require an IT infrastructure that is optimized for a level above its current needs to accommodate potentially more intensive processing or bigger data sources.

Automation, on the other hand, enables an organization to save on human resources. As listed in the table, the cost of training staff should be included in the initial costs of introducing ML methods, as the staff of statistical institutes is usually trained in classical statistics and may need appropriate training to use ML. This cost has to be sustained whether the application of ML is planned for small datasets or large datasets. However, the staff's underlying domain knowledge and statistical preparation should ensure that such training isnot too extensive; consequently, the transition training costs may not be high. But, as the field of ML is subject to rapid innovation and its application in official statistics is still new, the need for continuous learning cannot be neglected. For this reason, staff training is also an ongoing cost.

Once the fixed and ongoing costs of training are considered, automation should make it possible to save in terms of staff needed to execute operations. This should enable organizations to free up human resources for employment in other sectors of the statistical production cycle. In turn, the staff employed on ML procedures could then focus on aspects important for official statistics, such as explainability and the methods and inferential reproducibility of results.

Lastly, the adoption of ML algorithms opens new possibilities for data collection and data sources. From an IT point of view, acquiring big data sources presents the challenges that were illustrated before: expansion of storage space, both local and in the cloud, improvement of hardware, and so on. Additionally, acquisition costs must also be factored in, as big data sources are often held by private companies. Such costs may be either fixed or ongoing, depending on the agreements with data providers. In such cases, of course, it is advisable for an organization to try to obtain a test dataset to assess its usefulness for the current production before committing to an agreement. It is also worth reiterating that some big data sources can be freely accessed, for example, through web scraping or open data portals.

From the elements described above, some tests can be formulated to include the various aspects of cost effectiveness into the assessment of accuracy. First of all, the accuracy per unit cost metric described in Section 6.1 could be regarded as a cost-effectiveness test, useful for investigating the costs linked to an accuracy improvement deriving from the adoption of a new method. For this purpose, this test should include only the variable costs in its assessment, especially if used to compare an ML method with a traditional one, for which fixed costs have probably already been paid in previous years.

Another possible test focuses on the return on investment, which is useful to assess the fixed costs and the time needed to recoup the initial investment in ML. Two or more ML algorithms can be compared over a specific period of time (e.g., five years) to assess which offers more savings and whether such savings are enough to compensate for their introduction in the production process.

An ML algorithm should be chosen only if both tests produce satisfactory results, that is, if the algorithm is cost effective and the cumulative savings it guarantees are bigger than the net present value of the investment in ML.

The same ML and IT infrastructure can be—and usually is—shared between multiple ML procedures. This should happen as NSOs become more confident in ML methodologies and increase their adoption of ML. In this case, when the metrics are computed to evaluate the costs and savings of an ML implementation, fixed costs should be apportioned between the relevant algorithms.

6.5 Conclusions

The previous illustration of the potential costs of implementing ML should shed some light on the metric that was introduced at the beginning of the section, accuracy per unit cost. When this measure is computed, it can be convenient to differentiate between specific elements of the potential expenses, depending on the needs and the current state of the statistical organization. In other words, the accuracy per unit cost metric does not have a given single use, as it has to be considered in the context of each organization. For example, 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.

Lastly, if ML allows an organization to improve the accuracy of its estimates while saving some resources, the question of where best to redirect these resources should be investigated. Of course, this is another case-by-case question, and a general answer is impossible. In the context of official statistics, it is important to highlight that the experimental nature of the processes and the novelty of some of the techniques may call for additional quality measures and controls. Since the mission of official statistics programs is to produce transparent, accurate and accessible data, it may be worth spending some of the additional resources to maintain regular quality assurance and quality control operations for the processes involving ML. This would ensure greater transparency for data users and give data producers deeper insight into the technical aspects of ML.

7.0 Summary and recommendations

National statistical offices (NSOs) around the world are modernizing, and many are looking at modern statistical algorithms as a significant part of their modernization journey. Modern statistical algorithms have plenty to offer in terms of increased efficiency; potentially higher quality; and the ability to process new data sources, such as satellite images. The challenge comes from deciding when modern algorithms should be used to replace existing algorithms. Many modern algorithms were developed in a prediction context and are designed to minimize prediction error. However, most algorithms currently used in official statistics were developed to produce inferentially correct outputs. Comparing methods developed under these two paradigms is not easy.

The proposed Quality Framework for Statistical Algorithms (QF4SA) is a first attempt to lay down some groundwork to guide official statisticians in comparing algorithms (be they traditional or modern) in producing official statistics. The QF4SA's five dimensions are applicable to traditional and modern algorithms and provide food for thought to official statisticians when choosing between different algorithms. Based on the QF4SA, the working group proposes the following recommendations:

1) It is recommended that all five dimensions of the QF4SA be considered when deciding on the choice of an algorithm, particularly when choosing between traditional and machine learning (ML) algorithms.

2) Given that explainability is a major barrier to wide acceptance of ML algorithms, it is recommended that NSOs explore and use the methods outlined in the Explainability section to help users understand the relationship between input and output variables. Data users' understanding can help eliminate some of the black-box concerns associated with ML. This will contribute to increased acceptance of and trust in ML algorithms.

3) Ideally, NSOs should estimate the expected prediction error using methods such as cross-validation or other appropriate resampling methods. These methods are valid only if the training sets are generated from the data in the same way the data are generated from the population. This underlines the importance of properly constructed training sets. For instance, training data about only women should not be used to train a model to predict the income of men. Therefore, it is recommended that NSOs calculate the expected prediction error, as well as the prediction error, to protect against potentially poor-quality training data. In addition, it is recommended that high-quality training data (data that are representative of the population in question) be created when applying supervised ML algorithms. NSOs may want to leverage their sampling expertise to research the creation of high-quality training data.

4) Given the role that reproducibility plays in gaining the trust of data users, the QF4SA recommends that, as a minimum, NSOs take action to implement methods reproducibility. Inferential reproducibility should be carried out as well, when possible and desirable, limited to only the replication of the analysis using different but applicable algorithms and assumptions. We note that for inferential reproducibility, the results of a chosen method need only be corroborated by alternative algorithms or assumptions. They do not need to be the same. When the alternative algorithms or assumptions do not corroborate the original results, NSOs should ensure they understand why and determine whether the chosen method is warranted.

5) Timeliness is covered by most, if not all, existing quality frameworks. However, the timeliness dimension commonly used is defined as the time between the end of the reference period and the availability of the information sought. For certain processes leading to the production of statistical outputs, it is recognized that modern algorithms could lead to significantly shorter development and processing times, in comparison with traditional algorithms. Examples of these processes include industry and occupational coding and image processing. Therefore, the QF4SA recommends that development and processing time be added to the commonly used concept of timeliness.

6) A motivating factor of NSOs' modernization is cost effectiveness. By considering alternative data sources, NSOs want to reduce collection costs and respondent burden. For some alternative data sources (e.g., satellite images), modern algorithms are the only available way to process them. When evaluating the cost of potential algorithms, NSOs must consider fixed costs, as well as ongoing costs. Examples of fixed costs include establishing information technology (IT) infrastructure and retraining employees to work with the new infrastructure. We note that fixed costs can be amortized over time or across projects. Examples of ongoing costs include IT maintenance, cloud storage for the data, the

cost of acquiring the data and processing time. Processing time in particular could be significantly reduced under certain circumstances by using modern methods. Given these costs, the QF4SA recommends that NSOs consider two aspects in particular when considering cost effectiveness: cheaper operating costs and time to recoup fixed costs.

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.

Arrieta, B.A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115, doi:10.1016/j.inffus.2019.12.012.

Bengio, Y., and Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, 5, 1089–1105.

Bhatt, U., Xiang, A., Sharma, S., Weller,A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., and Eckersley, P. (2020). Explainable machine learning in deployment. arXiv:1909.06342.

Bickel, P.J., and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6), 1196–1217.

Biemer, P.P. (2010). Total survey error – Design, implementation, and evaluation. *Public Option Quarterly*, 74(5), 817–848, doi: 10.1093/poq/nfq058.

Borra, S., and Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics and Data Analysis*, 54, 2976–2989, doi: 10.1016/j.csda.2010.03.004.

DiCiccio,T., and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189–228.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530), 636–655, doi: 10.1080/01621459.2020.1762613.

European Union (2016). Regulation 2016/679: General Data Protection Regulation, Recital on Profiling, https://gdpr-info.eu/recitals/no-71/.

Eurostat (2014). *Handbook on Methodology of Modern Business Statistics*, CROS Portal,https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en.

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

Gleser, L.J (1996). Comment on 'Bootstrap confidence intervals' by DiCiccio and Efron, *Statistical Science*, 11(3), 219-221.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2014). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. arXiv:1309.6392.

Goodman, S., Fanelli, D., and Ioannidis, J. (2016).What does research reproducibility mean?*Science Translational Medicine*, 8(341), 341ps12, doi: 10.1126/scitranslmed.aaf5027.

Groves, R.M., and Lyberg, L. (2010). Total survey error – Past, present, and future. *Public Opinion Quarterly*, 74(5), 849–879, doi: 10.1093/poq/nfq065.

Hand, D.J. (2012). Assessing the performance of classification methods. *International Statistical Review*, 80(3), 400–414, doi: 10.1111/j.1751-5823.2012.00183.x.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* 2nd edition. Springer.

Japkowicz, N., and Shah, M. (2011).*Evaluating Learning Algorithms.*Cambridge University Press.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, 53, 3735–3745, doi: 10.1016/j.csda.2009.04.009.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91, 216–231, doi: 10.1016/j.patcog.2019.02.023.

Molnar, C. (2019).*Interpretable Machine Learning:A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

Platek, R., andSärndal, C.-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, 17(1), 1–20.

Prasath, V.B.S., Alfeilat, H.A.A., Hassanat, A.B.A., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B., and Salman, H.S.E. (2019). Effects of distance measure choice on K-nearest neighborclassifier performance: A review.*Big Data*,7(4), 221–248, doi: 10.1089/big.2018.0175.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353–360.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144, doi:10.1145/2939672.2939778.

Scholtus, S., and van Delden, A. (2020). *On the Accuracy of Estimators Based on a Binary Classifier*, discussion paper, CBS, The Hague/Heerlen.

Statistics Canada (2017). Quality Assurance Framework, 3rd edition,https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm.

Stodden, V., Seiler, J., and Ma, Z. (2018).An empirical analysis of journal policy effectiveness for computational reproducibility. *PNAS*,115(11), 2584–2589, doi: 10.1073/pnas.1708290115.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36, 111–147, doi: 10.1111/j.2517-6161.1974.tb00994.x.

Stothard, C. (2020). Evaluating machine learning classifiers: A review. Australian Bureau of Statistics, available upon request.

Szabo, L. (2019). Artificial intelligence is rushing into patient care—and could raise risks. *Scientific American*, December.

UNECE (United Nations Economic Commission for Europe) (2018). *The Use of Machine Learning in Official Statistics*,https://statswiki.unece.org/download/attachments/223150364/The%20use%20of%20machine%20learning%20in%20official%20statistics.pdf?version=1&modificationDate=1542811360675&api=v2.

United Nations (2014). Fundamental Principles of National Official Statistics, https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

United Nations (2019). *National Quality Assurance Frameworks Manual for Official Statistics*, https://unstats.un.org/unsd/methodology/dataquality/.

Vanwinckelen, G., and Blockeel, H. (2014). Look before you leap: Some insights into learner evaluation with cross-validation. *JMLR Workshop and Conference Proceedings*, 1, 3–19.

Vilone, G., and Longo, L. (2020). Explainable artificial intelligence: A systematic review. arXiv:2006.00093.

Wolter, K.M. (2007). *Introduction to Variance Estimation.*2nd edition. Springer.