A Quality Framework for Statistical Algorithms (QF4SA)

Executive Summary

To come.

1.0 Introduction

The aim of National Statistical Offices is to develop, produce, and disseminate statistics that can be considered as a reliable portrayal of reality (UNECE 2012). Quality is the degree to which a set of inherent characteristics of a statistic fulfils certain requirements (Eurostat 2014).These requirement are typically set out in a quality framework which is a set of procedures and systems that support quality assurance within an organisation and is meant to cover the statistical outputs, the processes by which they are produced, and the organisational 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 Nation's National Quality Assurance Framework (UNSD, 2019), Eurostat's Code of Practice of the European Statistical System (Eurostat, 2005) and Statistics Canada's Quality Assurance Framework (Statistics Canada, 2017).

Modern methods, such as machine learning (ML), are gaining popularity in the toolbox of the official statistician. In combination with modern hardware and software, these methods allow official statisticians to process new data sources such as text and images, to automate existing statistical processes and potentially to make inference without a sampling design. With this increased interest there is a need to consider a quality framework 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 in 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 square methods) to obtain the "estimated" surface.

Machine learning, 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 not worry about the surface plus noise models" [Ibid.]. Rather than searching for a hidden truth about the underlying phenomenon that generated the data or characteristics of the population, the primary aim is 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. β_0, β_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 machine learning 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. On the other hand, companies providing real-time traffic service for commuters would be more interested in predicting whether a certain route that the commuter is taking will be congested or not and this is the area of prediction where ML is specialised in. In applications where accurate predictions at the level of individuals are infeasible, ML methods may also see limited applicability. However, statistical methods can still deliver insight. For example, a statistical model such as logistic regression allows the assignment of significance to individual predictors when modeling 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 recommendation or search engine refinement is due to their ability to make predictions for individual cases. In the official statistics field, the use of ML is becoming increasingly popular in areas where such individual prediction tasks are needed. It can be in 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 coming about due to the acceptance of machine learners to use more complex models than traditional statisticians, which can lead to higher predictive accuracy.

ML is a relatively new tool in the official statistics field. While there is an increasing body of work on methodological aspects of ML, there has been less done on quality aspects. Commonly used and accepted quality concepts may require re-evaluation through ML perspectives. For example, the UN National Quality Assurance Framework writes "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". While this accuracy is often considered as how accurate statistical estimates that describes characteristics of the underlying population (e.g. unemployment rate estimate based on Labour Force Survey), “accuracy” for ML can also mean how accurate the predictions are for individual cases in an intermediate processing task as part of the entire production process in common application areas within statistical organisations. Also, unlike manual classification done by humans, ML methods are scalable but may require initial development and investment which affects cost-effectiveness and timeliness of the end-product in a different way compared to existing methods. The specificity of ML methods requires new quality dimensions (e.g. explanability and reproducibility) which are not considered in existing quality frameworks.

The goal of this document is to propose a 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 optimisation 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.

Under the QF4SA, we propose five quality dimensions; explainabiity, 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. The definitions of these dimensions are given below:

Explainability

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

Accuracy

Across several internationally accepted frameworks, slightly different definitions of accuracy are given. The definition proposed for 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 designs of the studies, experimental methods or analytical choices (inferential reproducibility).

Timeliness

Timeliness is defined as the time involved in producing a result from conceptualization, algorithm building, processing and production. Distinction should be made between timeliness in development and production. The former generally takes longer than the latter.

Cost effectiveness

Cost effectiveness is defined as the degree to which the results are effective in relation to its 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 (measured by the Mean Squared Error (MSE) or F1 scores for example) per unit cost. Note that the total cost of doing the job (scalability included), including fixed costs such as infrastructure, staff training etc. and ongoing costs such as production costs should be taken into account.

One could argue 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 reflect on 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 an output is produced from a statistical or a machine learning (ML) algorithm using its input features**. Throughout the document we use the term feature to represent individual independent variables that are inputs and 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 how input features are 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 as to why *X* produces 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 turns out to be an important factor. Explainability can therefore be pictured as a concept between the 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 liking inputs to outputs. For example, regression coefficients explain the direction and strength of the relationship between the 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), on the other hand, is straightforward.

While predictions from a single decision tree are explainable (e.g. the prediction of instance *i* is $\hat{\mathrm{y}}_{\mathrm{i}} = \mathrm{y}_{\mathrm{i}}$ 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 it.

2.2 Importance of explainability

Explainability is important to gain **trust** about ML algorithms from users. ML algorithms are often considered as a black-box. Understanding how an ML algorithm makes decisions can give more trust to

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

users as they can relate the behaviour of the ML algorithm with their prior knowledge and internal logic. We do note that explainability might be user-specific. For a statistical organisation using an ML, these users can include: statisticians who may not be familiar with ML methods and subject matter experts in the organisation; data providers in partner organisations; data users from the general public, academia and policy makers; as well as data scientists developing ML algorithms.

Understanding how algorithms make certain predictions can shed light to 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 performance of the algorithm itself (for ML developers).

While a high prediction accuracy of a ML algorithm indicates that the 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 for 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 for critical decision making which, in some circumstances, impact human life. Therefore by requiring some human intervention, explainability can serve as a **safeguard** for machines making correct decisions for the right reasons.

Explainability draws great attention to developing **fair, accountable, transparent** and **ethical** (FATE) artificial intelligence. In application areas where 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, etc.), it is important to ensure that such decisions are made in a fair and ethical way. For example, if a machine learning 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 as "unethical", hence should be corrected before deployment, regardless of how accurate the prediction of the algorithm is. ML algorithms are often considered as 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 performed. With increasing awareness that human subjects should be provided with an "explanation of the decision reached [through automated processing]" (GDPR Recital 71 on Profiling), national statistics offices, as a public agency, should be aware of these issues around the use of ML. For example, could the output of a machine learning process using many features identify unique individuals in a population?

2.3 Making predictions explainable

Explainable ML, or eXplainable artificial intelligence (XAI), is a recent but very active field of research. A multitude of methods, each with its own benefits and caveats, has been proposed to make predictions from “black-box” algorithms more explainable. Note that these methods do not make the machine learning algorithms more explainable directly. Instead, they make predicted results more explainable which sheds lights on the behaviour of the algorithm, hence 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 details are encouraged to refer to 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 show 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 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.

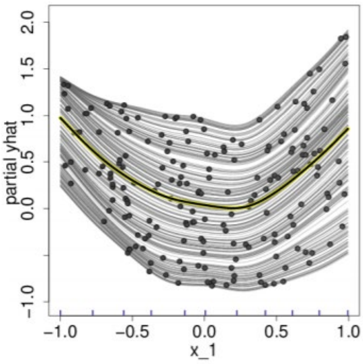

Figure 2.1. Example of Individual Conditional Expectation (ICE) Plot

(Source :Goldstein et. al., 2014)

Partial dependence plots (PDP) averages over all instances to show the overall marginal effect of a feature on the model prediction. While ICE and PDP are intuitive and easy to implement, they assume the feature of interest, plotted on x-axis, is uncorrelated with other features, which 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. shuffling its values 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 (LIME)** are an implementation of a surrogate model which is aimed to explain 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).

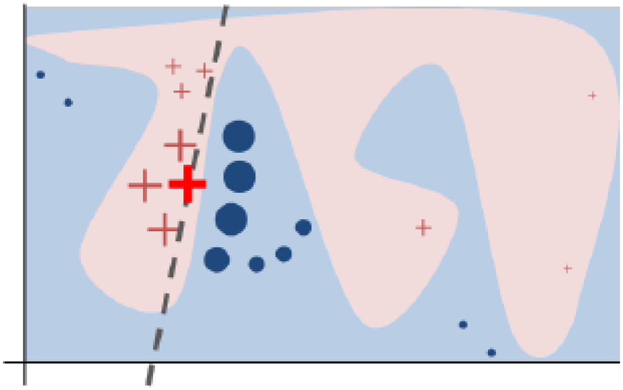

Figure 2.1. Example of **Local interpretable model-agnostic explanation (LIME)**

(Source: Ribeiro et. al., 2016)

The **Shapley value** is a measure for 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 find[2] data points in feature space that are aimed to serve as:

- **Counterfactual example**: a data point that is as close in feature space as possible to the instance of interest but with a different predefined outcome. For example, assume that a description of a

[2] In this note, we focus on describing what the data points of interest are but omit how to find those data points through optimization of loss functions.

work-related injury is “I cut my finger while chopping something on a wood board” was given 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**: an instance where 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 hence explainability.
- **Influential instance**: a data point in the training set that affects considerably 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.

Traditional statistical algorithms employ intuitive formulations which produce results that are innately explainable. Machine learning algorithms may produce a higher predictive accuracy than these traditional methods but due to their complexity, they are often considered as incomprehensible black-boxes which can hamper the use of ML in the statistical organizations. Therefore, as machine learning becomes more common in the production of official statistics, the QF4SA is recommending that if complex algorithms are used in any phase of the production of outputs, 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 machine learning, we note that there may be some confusion when discussing accuracy: the term *accuracy* is used for a specific performance indicator in classification and machine learning (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 machine learning.

We also note that depending on the variables involved, measures of accuracy could take on different forms. For continuous variables, the Mean Squared Error (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 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 area under the ROC curve (AUC).

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.

It can be stated that it relates 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 the estimate is. We note that the deviation may be structural (bias) or random (variance).

For every framework, qualifying comments are common. For instance, "Any factors which could impact on the validity of the information for users should be described in quality statements" (Australia) or "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" (Canada).

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. Such 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 National Statistical Offices includes the development, production, and dissemination of statistics that can be considered as a reliable portrayal of reality. To ensure that these statistics are of high quality, most National Statistical Offices have developed quality frameworks which 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 of them tailored to the particular estimation method being used and the situation (e.g. the type of data, the type of target parameter, etc.). Hence 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 measure is to provide quantification of the closeness of the estimate to the true value.

It is important to underline that the existing literature about the performance of an algorithm reflects the fact that to evaluate an estimator it is important to consider two different aspects (e. g., Hand 2012):

a) In choosing the estimator for the job, 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 in order to choose the most efficient one for the job.
b) After an estimator has been chosen, an assessment of how well the estimator can predict the true values of new data.

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 the uncertainty of it is required.

Thus, 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. Hence, it depends on the estimation method under study. Therefore, before going into detail on measures of accuracy, we first set the context of how machine learning algorithms are 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 machine learning 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 machine learning, the existence of a function linking the variables in the two sets is presumed,

$$Y = f(X).$$

A machine learning algorithm estimates the mapping function $\hat{f}$ from the input to the output. The goal is to approximate the mapping function so well that, when there is 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 hence the variables within), we differentiate the task as follows. If the space is continuous, that is, it consists of an infinite number of elements, then the task is called regression. On the other hand if the space is discrete, then the task is called classification. In less technical terms, if the output variable is quantitative/ numeric the learning task is called regression; if it is qualitative/categorical the learning task is called classification.

Regardless of whether the task is regression or classification, machine learning algorithms will attempt to learn the relationship between *X* and *Y* based solely on the available data observed in *S*. As such, machine learning algorithms can be much more flexible than traditional modeling methods as they do not tend to pre-suppose particular relationships between *X* and *Y*. Of course, one always has to be aware of the problem of overfitting as algorithms become more flexible, i. e. of the possibility that a learnt model fits very well on the observed data (perhaps because it even interpolates the data) but generalizes poorly to yet unseen data. Using pre-specified models with controls to avoid unnecessary complexity, for example very high dimensional polynomial terms, can reduce the danger of overfitting. Machine learning, however, is known to be more flexible in the sense that it often does not have such a restriction. Usually, the class of possible models is much larger than only polynomials which are susceptible to overfitting to the observed data. Regularization, stopping rules, and the evaluation of the learnt model on test data sets which have not been used during the learning process are schemes to deal with this potential problem and help to improve the generalizability of estimators or predictors. Thus, a machine learning model should be learnt 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 4.1). Note that Figure 4.1 is for illustrative purposes and is not suggesting an optimal number of validation or testing sets nor a ratio between the two.

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

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

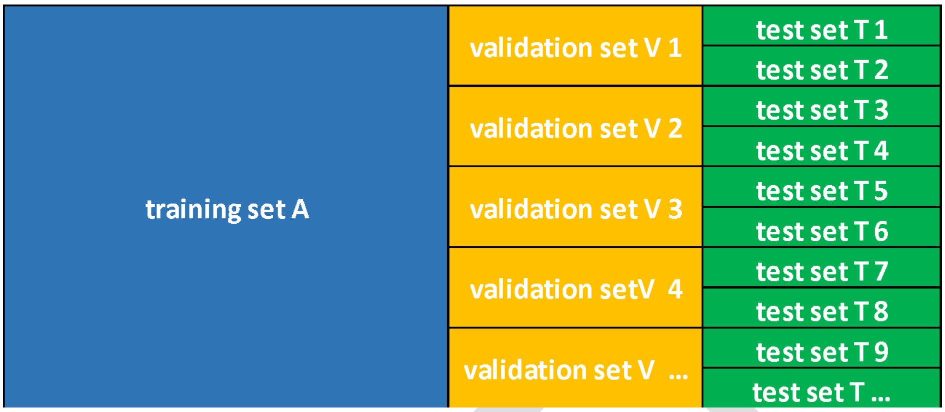

Figure 4.1: Training, Validation and Test sets

The set up described is Figure 4.1 is the best way to split the set *S*, but for various reasons practitioners may chose other ways. Often, when there are not multiple validation sets (in orange) bootstrapping or cross-validation on this single validation set is used to simulate the ideal situation where there are multiple validation sets. At times, due to 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 common version is to learn some models based on one training set (perhaps with cross-validation on it to specify some parameters) and to test it on only one test set (or do bootstrapping on this to approximate the situation above). By re-partitioning S into A and T multiple times, we have the opportunity to train and test the different algorithms or different parameters of the algorithms on multiple data sets, thus giving us their performance in choosing the most efficient algorithm/parameters.

3.3.2 General approach for assessment of accuracy

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

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

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

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

$$L\left(Y, \hat{f}(X)\right) = 1(Y \neq \hat{f}(X)) \quad .$$

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*. This error is often estimated by

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

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

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

The choice of which error to calculate depends on the situation at hand. If one is interested in the general performance of a machine learning algorithm, it is necessary to estimate *Err* which gives some protection from a poorly constructed training set. Furthermore, *Err* offers an impression about the robustness of an approach when the input data slightly varies. Fortunately, cross validation seems to estimate well Err (Hastie et al., 2009).

However, when a particular machine learning model (a predictor, an estimator) has already been learnt, it has been learnt based on a concrete training data set *A*, so *Err*$_A$ has to be calculated in order to get an impression about the future performance of this particular machine learning algorithm.

3.3.3 Variance

One common point of criticism of machine learning 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, which 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; 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 machine learning algorithms, results like these cannot, at this time, be produced for those approaches without making additional assumptions. We note that assumptions are also required in traditional methods.

In the context of both machine learning and traditional statistics, resampling methods like 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)–(G3) mentioned above. Wolter (2007) presents an introduction with a focus on the context of survey sampling, while some studies in the context of classification and regression are given by 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 being used. This is emphasized here because there are situations where, for example, the empirical bootstrap does not deliver suitable estimates (e. g., Bickel and Freedman 1981). However, their examples of bootstrap failures are unlikely to occur in official statistics. We next 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. It 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:

- 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 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 in Hastie et al. (2009, pg. 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*, but note that there does not exist an 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 calculate formally 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* is the total size of the set S. That is $N = \sum_k N_k$. If the folds are of equal size, that is N_k=*N/K*, then $\widehat{Var}_{CV}(pred)$ is equivalent $\widehat{Err}_{CV}$ except for a factor of *N* versus (*N-1)*. One could also look at the residuals, $\left(y_i - \hat{f}^{-k}(x_i)\right)$, to come up with something like empirical 95% prediction intervals, but again there is the limitation that the probability distribution of the cross-validation estimator is not known exactly. A critical discussion on cross-validation is given by Vanwinckelen and Blockeel (2014).

Bootstrap

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

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

Let T_b be the set of units which 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. (If bootstrapping is used at the validation step, use V_b instead of T_b as notation to avoid confusion). 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

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

One can also calculate the following quantities

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

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

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), p. 194, for some references). The proposed variance estimators borrow heavily from traditional statistical methods and the theoretical properties of them 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 (for example in R or Python) already provide them as standard procedures. Nevertheless, these techniques, which are surely necessary when using machine learning, 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 the evaluation of statistical algorithms or their results in ML

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 the design, collection, processing, and analysis of 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: That part of the difference between a population value and an estimate thereof, derived from a random sample, due to the fact that only a subset of the population has been enumerated (Eurostat).
- Non-sampling error: 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 total survey error accumulates all errors, which may arise in the sample design, data collection, processing and analysis of survey data, and it comprises both sampling and non sampling errors.

Technically, there are a number of measures that may be used to indicate accuracy through the definition of the 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 bias squared 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 predicting 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. In the univariate situation there are, however, well-known tests (like the Kolmogorov Smirnoff test) to 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 rules for imputed values, (if possible: the impact on the downstream task), and the accuracy (ideally also the variance) of the estimation of the target parameters which are eventually to estimate are further important figures. When measuring distribution accuracy, the Jensen-Shannon metric appears to be an appropriate metric as outlined in Prasath et al. (2019), because of its versatility to handle multivariate distributions with continuous and categorical variables.

3.4.3 When the focus is on cellwise predictive accuracy (G1)

Referring to the pilot studies undertaken within the UNECE HLG-MOS machine learning project and the literature (e. g., Japkowicz and Shah 2011, Pepe 2003, Stothard 2020, Hand 2012), measures commonly used to assess the success of machine learning algorithms are:

- In the case of regression, root mean squared error (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 and/or on macro levels, G measure, Matthews' correlation coefficient and awareness of the consequences of the different misclassifications (see fair, accountable, transparent and ethical (FATE) artificial intelligence).

The references mentioned above contain many more measures and discussion about them. A critical point in case of classification is for instance, how sensitive measures are to class imbalances (see, e. g., Luque et al 2019) or whether they need a pre-specified 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 discussion). Note that when these measures are estimated in a concrete task in order to evaluate how well the learnt predictor works, these numbers are only valid 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 implies that the accuracy of an ML model must be continuously monitored, and underlies 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, Steiler and Ma, 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 recognise 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), having followed the same experimental methods. This has previously been described as replicability.
- **Inferential reproducibility** is defined as the making of 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, recognising that it is not feasible for the official statistician to undertake new data collections to corroborate the initial findings, it is **NOT** proposed to adopt the Results reproducibility in official statistics.

In addition, consistent with the Fundamental Principles of Official Statistics, Methods reproducibility has been invariably embraced by National Statistical Offices and its adoption in the QF4SA when using statistical algorithms to produce official statistics is expected to receive overwhelming support.

On Inferential reproducibility, as there are generally multiple algorithms that may be brought to bear on the data analysis, there would be multiple ways to reanalyse the data. The official statistician, when deciding to use a particular algorithm, with a decided set of assumptions, for the analysis, has to be reasonably satisfied that the results from the chosen analysis can be corroborated from the 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 scores, of the difference between results obtained from the same data set, 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), and not whether that result is good.

An example to illustrate the difference between accuracy and reproducibility is as follows:

> 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 to predict Y. In that case, we would have poor prediction accuracy (high MSE for predictions of Y). As one will get the same inaccurate predictions on Y using the same model and assumptions, the analysis is Methods reproducible. Also, the result that X2 is irrelevant to predict 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 reproducible) or corroborating result (Inferential reproducible), and not whether that result is good.

4.2 Importance of reproducibility for official statistics

Reproducibility builds and enhances trust in official statistics. The third Fundamental Principle of the Fundamental Principle of Official Statistics, "Accountability and Transparency", adopted by the United Nations Statistical Commission in 1994, 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"

An underlying rationale for reproducibility is well articulated by Professor Gleser. In 1996, when commenting on their seminal paper on Bootstrap Confidence Intervals published in Statistical Science by DiCicio and Efron (1996), Professor 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".

In the academic world, to ensure the "First Law of Applied Statistics" is followed through, many journals have revised author guidelines to include data and code availability. For example, the prestigious journal, "*Science",* commencing February 11, 2011, requires:

> "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. Whilst there are many factors that contribute to the building of trust, an important one, as outlined also in the First Fundamental Principle, is impartiality which, in a large measure, can be demonstrated by transparency in the sources, methods and procedures in the compilation of official statistics. Such transparency will allow independent analysts or researchers to access the integrity of, or where possible, reproduce and verify, the published official statistics.

4.3 Demonstrating reproducibility

Those who develop statistical algorithms, e.g. methodologists, data scientists and analysts, to compile official statistics are encouraged to assess the Methods and Inferential reproducibility of their algorithms before adoption. Once the reproducibility dimensions of the algorithms has been confirmed, during the development stage, they can be put into production, and no further re-assessment would be considered needed**.**

Methods reproducibility refer to the provision of enough details about algorithms, assumptions and data so the same procedures could, in theory or in practice, be exactly repeated.

Documenting Methods reproducibility thus requires, at minimum, the sharing of analytical data sets (original raw or processed data), relevant metadata, analytical code, and related software. Because of confidentiality reasons, National Statistical Offices are generally not at liberty to share identifiable raw data for independent analysis. It is therefore proposed that the replication of the analyses be carried out in-house and by another individual, who should be at-arm-length from the original researcher to assess reproducibility.

For Inferential reproducibility, the methodologist 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, judgement should be applied when examining the results which 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 the methodologist explain why it is the case, e.g. due to an improvement in efficiency, and provide back-up of the explanation using statistical theory?

Finally, it is also proposed that the outcomes of the Methods and Inferential reproducibility be documented for longevity and, where possible, publish these as part of the Quality Declaration statement normally released together with the official statistical output.

Clearly, reproducibility of statistical algorithms is fundamental to uphold the trust of official statistical outputs. Whilst three types of reproducibility are recognised in this section, we propose NSOs adopt the Methods and Inferential Reproducibility to support their choice of statistical algorithms in producing NSO outputs.

5.0 Timeliness

5.1 Timeliness for statistical algorithms

Quality guidelines or frameworks by many NSOs (Statistics Canada, the Australian Bureau of Statistics, the Office for National Statistics and the OECD) define timeliness as the length of time between the reference period and the availability of information. The QF4SA is advocating the consideration of 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 the need for data and the release of the information to meet that need. With the increased use of big data, the speed at which machine learning algorithms can be trained and run could lead to significant improvement in timeliness. This is particularly true for processes which are typically done manually such as coding. Coding applications can be developed quickly using machine learning, particularly if past manually coded data can be used as training data. In addition to being able to be set up fairly quickly, once developed, machine learning algorithms are capable of processing vast amounts of data in a short period of time. In comparison to manual processes, machine learning algorithms could lead to significant savings in processing time.

5.2 Importance of timeliness

Official Statistics are only useful when they are relevant which means that they need to be available in a timely fashion. Economic indicators of a downturn in the economy are not relevant if they are only available 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 are considered as we feel that machine learning can offer some advantages over commonly used methods which can lead to improvements in the commonly used definition of timeliness.

5.3 Responsibility of ensuing timeliness

Typically the development and/or choice of statistical algorithms, be they machine learning or other commonly used methods, would be done or decided upon collectively by methodologists, data scientists and/or analysts. Depending on the problem at hand, the work should also involve informatics specialists and subject matter experts. The considerations of timeliness are most likely evaluated during the development phase of a project. However, during redesigns or continuous improvement opportunities, this aspect of timeliness should be considered.

5.4Aspects to consider

Clearly, measuring the time required to develop, set into place and use in production is straight forward. In the section, we list some aspects which need to be taken into account during the evaluation.

- Data cleansing required

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

- Informatics Infrastructure

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

- Preparation of training data

  Supervised machine learning algorithms require high quality training data and depending on the method , a large quantity of data is required. Existing data should be considered for training data if appropriate. Note that some traditional approaches also have a need for auxiliary data which can be time consuming to obtain. For instance non-machine learning coding algorithms typically need a data dictionary that is complete, accurate and up-to-date.

- Evaluation of data quality

  Many well established methods have processes for evaluating data quality. For instance, a well developed theory for variance due to imputation exists. However, new approaches may not have well defined processes to estimate quality indicators and may rely on resampling type algorithms (for example, cross validation or bootstrap) to evaluate quality. Depending on the algorithm, these resampling methods could take significant amounts of 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 machine learning 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 that have been expended to obtain them. Results in statistics are mainly measured in terms of accuracy: thus, it is natural to link cost effectiveness to the accuracy dimension and try to measure it under 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 machine learning, an organization may compare the accuracy of a machine learning 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 3.4).The same comparison could be made, of course, between two or more machine learning algorithms if the objective were to choose the most cost effective one, all other aspects considered. However, there may be some practical issues 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 are to be paid to launch the infrastructure for the new method. Machine learning, which can be heavily dependent on the underlying Information Technology (IT) infrastructure, may pose some challenges in this regard. In fact, fixed costs for machine learning mainly include the IT-related costs for the acquisition of new software and hardware and the costs of training the organization’s staff. They are different from the other category of costs that can be identified, the ongoing costs, which derive from a regular effort to keep the whole system running and up to date. In the following table a list of the possible costs involving a machine learning project is reported. It may be useful to note that also traditional methods present fixed costs, which however have been invested by NSOs over many years, so it is usually the case that no additional fixed expenditure is required for them.

Table 1. Potential additional fixed and ongoing costs for machine learning adoption

Cost component	Type	Purpose
IT infrastructure	*fixed*	Acquisition of necessary hardware and software

Cloud storage	*ongoing*	Acquisition of necessary on cloud storage space
IT maintenance	*ongoing*	Maintenance of IT infrastructure
Initial training of staff	*fixed*	Training of current staff on ML; may include hiring of new staff
Staff formation	*ongoing*	Keeping staff up to date regarding new ML developments
Data acquisition	*fixed / ongoing*	Acquisition and processing of new data sources
Quality assurance	*ongoing*	Quality assurance and control

The details of these components will be explained later in this section. For now it should be noted that machine learning 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 in the process, ML methods may present little additional costs. The elements shown in the table can be considered as a starting point for the comparison of ML and traditional methods; such comparison can be made by a) analysing whether the running costs for ML methods are cheaper than traditional ones or b) computing the number of years to recoup the investment needed on the extra elements outlined in the table.

6.2 Advantages of cost effectiveness

The last decade has seen an explosion in data production, due to improvements in speed of computer processing 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 face the challenge to meet the required high-quality standards of official statistics with the resources that are made available to them. Cost effectiveness, indeed, is an aspect that has guided many statistical institutes in recent years: the European Statistics Code of Practice, for example, dedicates its principle n. 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 National Statistical Offices from a survey-centered data production to processes involving administrative and innovative sources of data. The introduction of machine learning can be seen as a further step in this evolution.

6.3 Organizational considerations

Machine learning in official statistics is still a field under investigation, although it has shown promising results. However, organizations are different from each other regarding their available budget and their statistical production, so the convenience of introducing machine learning into the current production has to be looked at on a case-by-case basis. If an organization is new to machine learning 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 starting costs and evaluate them against its budget, against the cost of the current production and the expected accuracy improvement. The 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 to the future savings that machine learning would grant. As such, fixed costs could actually be considered as 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 towards machine learning than others. It is possible that an organization is involved in many projects that can easily – and beneficially –adopt a machine learning approach, while another organization may have too few of such projects, so in this case the initial investment would be harder to justify.

6.4 The potential costs of machine learning

As it can be seen from table 1, a big part of the costs linked to the adoption of machine learning are IT-related and staff-related costs. In order to illustrate them it is convenient to introduce two of the main advantages of machine learning methods: scalability and automation.

The former implies that a procedure may 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, machine learning methods per se do not necessarily require any additional effort in terms of computations or resources. However, when used in conjunction with big data, they can quickly become computationally intensive. Machine learning 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 (CPUs, GPUs, storage space) before it can be employed for computational intensive operations or large data sets. Furthermore, IT costs should also include the resources needed for cloud storage and on-cloud computations, which are usually ongoing costs. In conclusion, when planning for the introduction of machine learning in a statistical process, an organization could require an IT infrastructure that is optimized for a level above its current needs to accommodate a potentially more intensive processing or bigger data sources.

Automation, on the other hand, lets an organization save human resources. As listed in the table, the cost of training staff should be included in the initial costs of the introduction of machine learning methods, as usually the staff of statistical institutes is trained on classical statistics and may need appropriate training for the use of machine learning. This cost has to be sustained whether the application of machine learning is planned for small datasets or large datasets. The underlying domain knowledge and statistical preparation of the staff, however, should ensure that such training will not be too extensive; consequently, the starting training costs may not be expensive. On the other hand, as the field of machine learning 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, the training of the staff also represents an ongoing cost.

Once the fixed and the ongoing costs of training are considered, automation should make it possible to obtain savings in terms of staff needed to execute operations. This should let the organizations be able to free up human resources to be employed in other sectors of the statistical production cycle. In turn, the staff employed on machine learning procedures could be able to focus on aspects that are important for official statistics, like explainability and methods and inferential reproducibility of results.

Lastly, the adoption of machine learning algorithms opens new possibilities for data collection and data sources. Acquisition of big data sources, from an IT point of view, presents the challenges that were illustrated before: expansion of storage space, both local and on cloud, improvement of hardware and so on. Additionally, acquisition costs must also need to 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 should be advisable for an organization to try to obtain a test set of the data in order 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 to investigate the costs linked to an improvement of accuracy deriving from the adoption of a new method. For this purpose, this test should only include the variable costs in its assessment, especially if used to compare a ML method to a traditional one, for which fixed costs probably have already been paid during the previous years.

Another possible test focuses on the return of investment, which on the other hand is useful to assess the fixed costs and the period of time that is needed to recoup the initial investment in ML. Indeed, two or more ML algorithms can be compared over a specific period of time (e.g. 5 years) to assess which one offers more savings and if such savings are enough to compensate for their introduction in the production process.

After these tests have been applied, a ML algorithm should only be chosen if both tests are satisfied, 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.

It is possible, and it is usually the case, that the same ML/IT infrastructure is shared between multiple ML procedures. This should happen as a NSO becomes more confident in ML methodologies and increases its adoption of ML. In this case, when computing the metrics to evaluate the costs and savings of a ML implementation, the fixed costs should be apportioned between the relevant algorithms.

6.5 Conclusions

The previous illustration of the potential costs of a machine learning implementation should shed some light on the metric that was introduced at the beginning of the section, the accuracy per unit cost. When computing this measure it can be convenient to differentiate between specific elements of the potential expenses, depending on the needs and the current state of a statistical organization. In other words, there cannot be a unique use of the accuracy per unit cost metric, as it has to be considered in the context of each organization. For example, decomposing it into the different components of cost is useful to better assess the potential savings and accuracy improvements against the future ongoing costs. This would also help to get an estimate of the time that would be needed to recoup the initial investment.

Finally, in the case that machine learning would allow an organization to improve the accuracy of its estimates while saving some resources, the question of the best destination of these resources should be investigated. Of course, this is another case-by-case question and a general answer would be impossible. In the context of official statistics, it can be 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 control. Since the mission of official statistics concerns the production of transparent, accurate and accessible data, it may be worth spending some of the additional resources to maintain regular operations of quality assurance and quality control for the processes involving machine learning. This would ensure greater transparency for the users of the data and a deeper insight on the technical aspects of machine learning for the data producers.

7.0 Summary and recommendations

NSOs around the world are modernizing and many are looking at modern statistical algorithms to play 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 sources of data such as satellite images. The challenge comes from deciding on when modern algorithms should be used or should replace existing algorithms. Many modern algorithms have been developed under a prediction context and are designed to minimize prediction error. However, most algorithms currently used in official statistics have been 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 five dimensions are applicable for traditional and modern algorithms and provide food for thought to official statisticians in choosing between different algorithms. Based on the QF4SA, the working group is proposing the following recommendations:

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

2) Understanding that explainability is a major barrier for wide acceptance of ML algorithms, it is recommended that NSOs explore/use the methods outlined in the Explainability chapter to help users understand the relationship between input and output variables. This understand by data users can help eliminate some of the black box concerns associated with ML. This will contribute to an increased acceptance and trust in ML algorithms.

3) Ideally, NSOs should estimate the expected predication error using methods such as cross validation or other appropriate resampling methods. The use of these methods is valid only if the training sets are generated from the data in the same fashion as the data are generated from the population. This underlines the importance of properly constructed training sets. For instance, training data containing only females should not be used to train a model to predict male incomes. It is, therefore, recommended that NSOs use high quality training data when applying prediction algorithms, which will also facilitate the estimation of expected prediction errors. In this context high quality training data means data that is representative of the population in question.

4) Recognizing the role that reproducibility plays in gaining the trust of data users, the QF4SA is recommending that, as a minimum, NSOs will take action to give effect to the implementation of Methods reproducibility. As well, where it is possible and desirable to do so, Inferential reproducibility, limited to only the replication of the analysis using different but applicable algorithms and assumptions, should be carried out as well. 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, the NSO should ensure that it understands why it is so and convince itself that the method chosen is warranted.

5) Timeliness is a dimension which 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 statistics outputs, it is recognized that modern algorithms could lead to significantly shorter development and processing times in comparison to traditional algorithms. Examples of these processes include industry and occupational coding and image processing. Therefore, the QF4SA is recommending that development and processing time be added to the commonly used concept of timeliness

6) A motivating factor of the modernization of NSOs is cost effectiveness. By considering alternative data sources, NSOs are looking to reduce collection costs and respondent burden. For some alternative data sources, satellite images for example, modern algorithms are the only choice available 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 the establishment of IT infrastructure and retraining of employees to work in the new infrastructure. We note that the fixed costs can be amortized over time or across projects. Examples of ongoing costs include IT maintenance, cloud storage for the data, cost of acquisition of the data and processing time. Processing time in particular could be significantly reduced under certain circumstances by using modern methods. With these costs in mind, the QF4SA is recommending NSOs consider two aspects in particular when considering cost effectiveness: Cheaper operating costs and Time to recoup fixed costs.

References

Australian Bureau of Statistics (2005). Data Quality Framework, Australian Bureau of Statistics, 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.

Begley C. and Ioannidis, J. (2015).Reproducibility in science: Improving the standard for basic and preclinical research.*Circulation Research*, 116(1),116–126, doi: 10.1161/CIRCRESAHA.114.303819.

Bengio, Y. and Grandvalent, 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

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

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

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.

Hanson, B., Sugden, A. and Alberts, B. (2011). Making data maximally available. *Science,* 331(6018), 649, doi: 10.1126/science.1203354.

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.

Matthews, S., Patak, Z. and Picard, F. (2020). Time series modelling to produce economic indicators in (near) real-time. Internal Statistics Canada document.

Molnar (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.

Petkovic (2020) AI and trust: explainability, transparency. Ethical implications of AI and AI Tools Lab, Frankfurt Big Data Lab, Goethe University.

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

Prasath, V.B.S., Alfeilate, H.A.A., Hassanate, A.B.A., Lasassmehe, O., Tarawnehf, A.S., Alhasanatg, M.B. and Salmane, 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, 353–60.

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.

Statistics Canada (2017). Quality Assurance Framework, 3rd edition, Statistics Canada,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 2019.

United Nations (2012). Guidelines for the template for a generic national quality assurance, https://unstats.un.org/unsd/statcom/doc12/BG-NQAF.pdf.

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.