Improving the output quality of official statistics based on machine learning algorithms

Meertens, Q.A.; Diks, C.G.H.; van den Herik, H.J.; Takes, F.W.

DOI
10.2478/jos-2022-0023

Publication date
2022

Document Version
Final published version

Published in
Journal of Official Statistics

sciendo

Journal of Official Statistics, Vol. 38, No. 2, 2022, pp. 485–508, http://dx.doi.org/10.2478/JOS-2022-0023

Improving the Output Quality of Official Statistics Based on Machine Learning Algorithms

Q.A. Meertens[1], *C.G.H. Diks*[2], *H.J. van den Herik*[3], *and F.W. Takes*[3]

National statistical institutes currently investigate how to improve the output quality of official statistics based on machine learning algorithms. A key issue is concept drift, that is, when the joint distribution of independent variables and a dependent (categorical) variable changes over time. Under concept drift, a statistical model requires regular updating to prevent it from becoming biased. However, updating a model asks for additional data, which are not always available. An alternative is to reduce the bias by means of bias correction methods. In the article, we focus on estimating the proportion (base rate) of a category of interest and we compare two popular bias correction methods: the misclassification estimator and the calibration estimator. For prior probability shift (a specific type of concept drift), we investigate the two methods analytically as well as numerically. Our analytical results are expressions for the bias and variance of both methods. As numerical result, we present a decision boundary for the relative performance of the two methods. Our results provide a better understanding of the effect of prior probability shift on output quality. Consequently, we may recommend a novel approach on how to use machine learning algorithms in the context of official statistics.

Key words: Output quality; concept drift; prior probability shift; misclassification bias.

1. Introduction

Recently, the demand for readily available official statistics with a high data resolution has increased rapidly. In the pandemic context, an eminent example is the demand for accurate statistics on the number of deceased persons and the causes of their death. For policy makers, it is crucial that such statistics are available swiftly and at a high frequency. Also in other contexts, national statistical institutes (NSIs) experience an increase in the demand for new, more frequent, and more detailed official statistics (Braaksma and Zeelenberg 2015).

1.1. A Paradigm Shift in Official Statistics

Traditionally, NSIs adopt strict quality frameworks when collecting data and producing official statistics. Two important quality frameworks include the OECD quality

[1] Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, the Netherlands. Email: q.a.meertens@uva.nl
[2] University of Amsterdam, Center for Nonlinear Dynamics in Economics and Finance, Roetersstraat 11, 1018 WB Amsterdam, the Netherlands. Email: c.g.h.diks@uva.nl
[3] Leiden University, Niels Bohrweg 1, 2333 CA Leiden the Netherlands. Emails: h.j.van.den.herik@law.leidenuniv.nl and f.w.takes@liacs.leidenuniv.nl
Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors would like to thank Sander Scholtus for his useful comments on an earlier version of the manuscript.

framework (OECD 2011) and the Regulation on European Statistics (European Commission 2009) translated into the European Statistics Code of Practice (Eurostat 2017). With such quality frameworks as a foundation, NSIs could be categorised as part of the *data modelling culture* as defined by Breiman (2001). Within that culture, a stochastic data model is the basis of any further statistical analysis.

Currently, a paradigm shift from a *data modelling culture* to an *algorithmic modelling culture*, as envisioned by Breiman (2001), is taking place in the field of official statistics (De Broe et al. 2020). Indeed, many NSIs have initiated experiments with supervised machine learning algorithms with the purpose of producing new or improved official statistics. Beck et al. (2018) provide a list of 136 machine learning projects at NSIs in 25 countries. In many projects, machine learning was used for classification (78) or for imputation (22). The results of these machine learning projects are promising.

Moreover, the quality of official statistics has a broad interpretation nowadays. Statistical quality (e.g., statistical accuracy) as well as information quality (e.g., data resolution, temporal relevance and chronology of data and goal) should be appreciated when evaluating official statistics (Kenett and Shmueli 2016). With such a broad quality framework in mind, official statistics based on algorithmic modelling are quite promising. A particularly appealing aspect of information quality in the context of machine learning is *construct operationalisation*. It questions whether the measured variables are of interest to the study goal (Kenett and Shmueli 2016).

Many of the projects presented by Beck et al. (2018) that use machine learning for classification, were intended to measure variables that could only be measured by employing machine learning algorithms. A typical example is identifying small innovative companies (too small to be included in traditional surveys), using data from their websites (Daas and Van der Doef 2020). Therefore, we aim for the best of Breiman's two cultures, complementing statistical quality with information quality.

1.2. Statistical Output Quality and Misclassification Bias

In this article, we consider classification algorithms (having high information quality) and investigate how to maximise their statistical output quality. We measure statistical output quality by the mean squared error (MSE) of the aggregated output as estimator of the desired quantity (Buelens et al. 2016). As a first step, we consider the simplest type of statistical output produced by NSIs, namely the proportion (relative occurrence) of a category of interest within a target population of objects. It corresponds to binary classification problems in machine learning. The proportion that we intend to estimate is referred to as the *base rate* and is denoted by a. Although this type of statistical output is rather uncomplicated, we stress that it occurs in a wide variety of applications, also outside official statistics. Examples include counting solar panels (Curier et al. 2018), estimating the relative occurrence of small innovative companies (Daas and Van der Doef 2020), measuring deforestation and other applications of land cover mapping (Costa et al. 2018), and predicting election outcomes based on sentiment analysis (O'Connor et al. 2010). Moreover, results for estimating the base rate generalise to more complicated statistical output. An important example is aggregating a numerical variable over the subpopulations obtained from the classification algorithm (Van Delden et al. 2016). The result of

aggregating the outcome of a classification algorithm will be referred to as a *classifier-based statistic*.

The definition of statistical output quality, that is, the accuracy of the algorithm measured by the MSE of its aggregated output as estimator of α, differs conceptually from that of algorithmic accuracy as used in the machine learning literature. The difference is that algorithmic accuracy is measured at the level of *individual* data points while statistical accuracy is measured at the *population* level (Forman 2005; Scholtus and Van Delden 2020). In fact, classification algorithms that have high algorithmic accuracy might still produce highly biased statistical output. The statistical bias in estimating a that can be attributed to the algorithm's incorrect classifications is referred to as *misclassification bias*.

Misclassification bias occurs in general when dealing with measurement error in categorical data. Consequently, it occurs in applications outside machine learning as well. It can be argued that any categorical variable (either statistically modelled or directly measured) containing incorrect classifications will result in misclassification bias when aggregated. This observation was first made by Bross (1954). Since then, misclassification bias has been neglected or overlooked by a large part of the statistical community (Schwartz 1985; González et al. 2017).

Fortunately, after several decades of scientific research, a rich body of statistical literature on misclassification bias is currently available. Tenenbein (1970) and Kuha and Skinner (1997) provide significant contributions to the literature on misclassification bias. A relatively recent overview is provided by Buonaccorsi (2010). Misclassification bias can be reduced significantly, if some form of extra information is available. In the broad context of categorical data analysis, this extra information can be, for instance, replicate values, validation data, or instrumental variables (Buonaccorsi 2010). Although such extra information might not always be available, in general, it is available in the context of supervised machine learning that we are considering here. In the context of machine learning, the extra information is validation data. Such data are often used for model selection, training and testing. We will use the test set as validation data to estimate error rates, and thus to correct misclassification bias. Below, we will argue why correcting misclassification bias in the context of official statistics is challenging.

1.3. Prior Probability Shift and Our Problem Statement

When machine learning projects at NSIs are initiated, the test set often is a random sample from the target population. The setup then corresponds to the *double sampling scheme* introduced by Tenenbein (1970). Among the correction methods discussed by Buonaccorsi (2010), the one proposed by Tenenbein (1970), referred to as the *calibration estimator* in this article, outperforms all the others in terms of MSE (Kloos et al. 2020).

However, correcting misclassification bias in the production process of official statistics is more challenging. In that situation, a statistical model is often estimated once and then applied for a longer period of time without updating the model parameters. This is common in the context of supervised machine learning, because otherwise new data have to be annotated manually in each publication period leading to high production costs. Still, the key issue remains that both the data distribution and the relation between the dependent and independent variables might change over time, causing the outcome of the model to

become biased. In the machine learning literature, this issue is known as *concept drift*. It has been investigated in stream learning and online learning for several decades (see Widmer and Kubat 1996), dating back at least to the work on incremental learning in the 1980s (Schlimmer and Granger 1986). The term *concept* was originally used for a set of Boolean-valued functions (Helmbold and Long 1994). Currently, it has a statistical interpretation that is more closely related to our setting. Nowadays, Webb et al. (2016) state that the term *concept* refers to the joint distribution $\mathbb{P}(Y, X)$, with class labels (dependent variable) Y and features (independent variables) X, as proposed by Gama et al. (2014). Allowing such a joint distribution to depend on a time parameter t, concept drift in the setting of supervised learning means that $\mathbb{P}_{t_1}(Y, X) \neq \mathbb{P}_{t_2}(Y, X)$, for $t_1 \neq t_2$. The effect of concept drift is that misclassification bias might increase even further over time.

In this article, we aim to prove theoretically which of two popular correction methods discussed by Buonaccorsi (2010), that is, the *misclassification estimator* and the *calibration estimator*, reduces the MSE of the base rate a most. The most restrictive type of concept drift is known as *prior probability shift* (Moreno-Torres et al. 2012). We investigate prior probability shift as a first step in understanding the effect of concept drift on the output quality of official statistics based on machine learning algorithms. In summary, our problem statement reads: how to reduce misclassification bias of classifier-based statistics that are affected by prior probability shift?

1.4. Three Contributions and an Overview of the Article

Our article focuses on the production process of official statistics (where concept drift arises) and contains three contributions.

The first contribution consists of analytical derivations of the bias and variance of the misclassification estimator and calibration estimator as estimators of the base rate α. We show that the calibration estimator is no longer unbiased when it is affected by prior probability shift. Moreover, we provide a sharp lower bound for the absolute value of its bias. Consequently, the conclusions drawn by Kloos et al. (2020) fail to hold when prior probability shift arises.

The second contribution is that we show how the optimal choice for a correction method depends on three parameters, (1) the model accuracy, (2) the class distribution (or class imbalance) and (3) the size of the test set. By visualising how the decision boundary depends on these three parameters, the article contributes to *concept drift understanding* Lu et al. (2019). It complements *concept drift quantification* (Goldenberg and Webb 2019) and *concept drift adaptation* (Gama et al. 2014).

Our third contribution is a practical recommendation on how to implement classification algorithms in the production process of official statistics. The recommended approach is based on the decision boundary as a function of the three parameters (1)–(3).

The remainder of the article is organised as follows. In Section 2, we introduce the notation, enumerate our assumptions, and define the two popular methods to reduce misclassification bias. Section 3 contains our analytical results on the bias and variance of the misclassification estimator and the calibration estimator, assuming prior probability shift. Subsequently, in Section 4, we numerically investigate the location and shape of the decision boundary as a function of the three parameters (1)–(3). Based on our numerical

results, we provide a practical recommendation on how to implement classification algorithms in the production process of official statistics (Subsection 4.4). In Section 5, we present our conclusions and suggest four promising directions for future research.

2. Methods

In the context of official statistics, the convention is to use the MSE to evaluate statistical output quality, also when using statistical models (Buelens et al. 2016). Within that context, our article focuses on prior probability shift. We recall that our problem statement reads: how to reduce misclassification bias of classifier-based statistics that are affected by prior probability shift? Admittedly, the answer depends on the assumptions made. Therefore, we will describe our assumptions and their implications carefully in this section.

In Subsection 2.1, we will introduce the notation and make three general assumptions. Subsequently, we provide a precise definition of misclassification bias. In Subsection 2.2, we will then provide two methods to reduce misclassification bias. Next, we distinguish two cases. In the first case, described in Subsection 2.3, we assume that manually annotated data are available in each publication period (month, quarter, or year). The assumption corresponds to the double sampling scheme (Tenenbein 1970). In the second case, described in Subsection 2.4, we drop the assumption on frequent availability of manually annotated data. Instead, we consider prior probability shift (Moreno-Torres et al. 2012).

2.1. Notation and Three General Assumptions

Consider a population I of N objects (households, enterprises, aerial images, company websites, or other text documents) and some target classification s_i (stratum) for each object $i \in I$. In this article, we restrict ourselves to dichotomous categorical variables, that is, $s_i \in \{0,1\}$, where category 1 indicates the category of interest. A compelling example is the use of aerial images of rooftops to identify houses (the objects indexed by i) with solar panels ($s_i = 1$) (Curier et al. 2018). We will now provide three general assumptions.

The first general assumption (G1). We assume that there is some (possibly time consuming or otherwise expensive) way to retrieve the true category s_i for each $i \in I$. In the example of identifying houses with solar panels, the true category could be obtained by manually inspecting the aerial images and annotating them with a label indicating whether the image contains a solar panel.

The second general assumption (G2). We assume that background variables or other features in the data contain sufficient information to estimate s_i accurately, that is, with few classification errors. More specifically, we assume that a machine learning model can be trained of which the probabilities of correct classification are strictly larger than 0.5 for each of the two classes. To that end, we proceed as follows. We draw a small random sample from the population and determine the true category s_i for the objects in the sample, see general assumption (G1). Then, the obtained data are split at random into two sets. The first set is used for model selection and training, that is, to select a machine learning model and to estimate its parameters. The second set is used to estimate the out-of-sample prediction error of the model. It is referred to as the *test set* and it is denoted by

$I_{\text{test}} \subset I$. The number of observations in the test set is denoted by n, which is very small compared to N.

The third general assumption (G3). The machine learning model is then used to produce an estimate $\hat{s}_i$ of the true category to which object i belongs. We assume that the probabilities of correct and incorrect classifications of the model depend on i, but only through the true value of s_i. More precisely, we let p_{ab} be the probability that $\hat{s}_i = b$ given $s_i = a$, for $a, b \in \{0, 1\}$. This assumption specifies the *classification error model* as introduced by Bross (1954), following the notation in Van Delden et al. (2016).

In addition, we adopt the notation a_i, which is a 2-vector equal to (1,0) if $s_i = 1$ and (0,1) if $s_i = 0$. The sum of all a_i is the 2-vector of counts v. We then define the 2-vector $a = v/N$ and denote its first component by α. The scalar α is called the *base rate*. The vector a_i is estimated by $\hat{a}_i$, which is obtained by replacing s_i by $\hat{s}_i$ in the definition of a_i. We obtain the estimate $\hat{v}$ of v as the sum of all $\hat{a}_i$. The resulting estimates of the vector α and the base rate α are denoted by $\hat{\alpha}_m$ and $\hat{\alpha}_m$ to indicate that they are produced by the machine learning algorithm. It is immediate that $\mathbb{E}[\hat{\alpha}_m] = P^T\alpha$, where P is the matrix given by

$$P = \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix}. \tag{1}$$

In general, $P^T\alpha \neq \alpha$, which indicates that $\hat{\alpha}_m$ is a biased estimator of the base rate α. The statistical bias of $\hat{\alpha}_m$ as estimator of the base rate α is referred to as *misclassification bias*.

2.2. *Two Methods to Reduce Misclassifiction Bias*

A wide range of methods that reduce misclassification bias is available, see Chapter 2 in Buonaccorsi (2010). Five of these correction methods were compared by Kloos et al. (2020), who conclude that two correction methods are most promising. The first correction method is the *misclassification estimator* $\hat{\alpha}_p$. It is defined as the first component of the 2-vector

$$\hat{\alpha}_p = (\hat{P}^T)^{-1}\hat{\alpha}_m, \tag{2}$$

in which $\hat{P}$ is the row-normalised confusion matrix obtained from the test set, that is,

$$\hat{P} = \begin{pmatrix} \hat{p}_{11} & \hat{p}_{10} \\ \hat{p}_{01} & \hat{p}_{00} \end{pmatrix} = \begin{pmatrix} \dfrac{n_{11}}{n_{11}+n_{10}} & \dfrac{n_{10}}{n_{11}+n_{10}} \\ \dfrac{n_{01}}{n_{01}+n_{00}} & \dfrac{n_{00}}{n_{01}+n_{00}} \end{pmatrix}, \tag{3}$$

in which n_{ab} denotes the number of objects i in the test set for which $s_i = a$ and $\hat{s}_i = b$.

The second correction method is the *calibration estimator* $\hat{\alpha}_c$. It is defined as the first component of the 2-vector

$$\hat{\alpha}_c = \hat{C}\hat{\alpha}_m, \tag{4}$$

in which $\hat{C}$ is the column-normalised confusion matrix obtained from the test set, that is,

$$\hat{C} = \begin{pmatrix} \hat{c}_{11} & \hat{c}_{10} \\ \hat{c}_{01} & \hat{c}_{00} \end{pmatrix} = \begin{pmatrix} \dfrac{n_{11}}{n_{11}+n_{01}} & \dfrac{n_{10}}{n_{10}+n_{00}} \\ \dfrac{n_{01}}{n_{11}+n_{01}} & \dfrac{n_{00}}{n_{10}+n_{00}} \end{pmatrix}. \tag{5}$$

The entries of the matrix are referred to as the estimated *calibration probabilities*.

2.3. The Double Sampling Scheme

In the context of categorical data analysis, Tenenbein (1970) proposed the *double sampling scheme* to improve categorical data that suffer from measurement error. His proposal is to obtain accurate measurements for a small random subset of the data (as such measurements will be expensive to obtain) and to correct population estimates using the calibration estimator from Equation (4).

The context of machine learning corresponds to the context of categorical data analysis as follows: the algorithmic predictions are the categorical data that suffer from measurement error and the accurate (but expensive) measurements correspond to the manually annotated data. In the context of machine learning, the double sampling scheme corresponds to the following (Double sampling) assumption (D1): the test set I_{test} is a random sample from the population I. If we assume (G1)-(G3) as well as (D1), then the MSE of the calibration estimator $\hat{\alpha}_c$ is always (for any model applied to any data set) smaller than that of the misclassification estimator $\hat{\alpha}_p$ (Kloos et al. 2020).

2.4. Prior Probability Shift

Official statistics on a particular social or economic indicator are often produced for a long period of time and are published periodically (monthly, quarterly, or annually). The output quality of indicators produced by NSIs is required to be high. A key issue in using classification algorithms in the production process of official statistics is that the target population I changes over time, including the background variables x_i and the base rate α. Therefore, the test set drawn at random from the population at one publication period cannot be viewed as a random sample from the population at the next publication period. A first solution would be to draw a new test set from the population (and then manually annotate the data) at each publication period for as long as the statistical indicator is produced. It corresponds to assumption (D1), see Subsection 2.3. However, due to cost constraints, such frequent data annotation is infeasible in practice. Thus, we must relax assumption (D1) and subsequently investigate the results achieved by Kloos et al. (2020) in the context of official statistics.

Henceforth, we will retain the three general assumptions (G1)-(G3), and we will replace assumption (D1) by *prior probability shift*. We follow the definition by Moreno-Torres et al. (2012), which states that prior probability shift is captured by the following two assumptions:

(P1) the class s_i causally determines the features x_i that are used to model $\hat{s}_i$, and (P2) the causal relation does not change between (at least) two consecutive months or quarters.

Assumption (P1) seems reasonable in many applications, three of which we include. Moreno-Torres et al. (2012) specifically mention medical diagnosis, where the disease

causally determines the symptoms. A second application is sentiment analysis (see O'Connor et al. 2010), where the writer's sentiment causally determines the words that the writer chooses. A third application is land cover mapping (Costa et al. 2018), where the mapped object causally determines the pixel values in the image. Assumption (P2) implies that $\mathbb{P}(\hat{s}_i|s_i)$ does not change between consecutive months or quarters. However, the base a is allowed to change, in contrast to the setup captured by assumption (D1), that is, the double sampling scheme. The interpretation of assumption (P2) should be that the change in the causal relation between consecutive months or quarters is sufficiently small to neglect it. Assumption (P2) should not be repeated indefinitely, but only for a limited period of time, say for a year.

In the setting of prior probability shift, we consider two different moments in time, say t_1 and t_2, with $t_1 < t_2$. The model parameters as introduced in Subsection 2.3 refer to their value at time t_1. For the value of the parameters at time t_2 we add an apostrophe to the notation. For example, α refers to the base rate at time t_1 and a' refers to the base rate at time t_2.

The test set $I_{\text{test}} \subset I$ has been obtained as a random sample from the target population I at time t_1. The aim is to estimate the base rate α' at time t_2 within population I'. We use predictions $\widehat{s_{i'}}'$ for $i' \in I'$ produced by the same model as the one trained at time t_1. In particular, it holds that $\hat{s}_i{}' = \hat{s}_i$ for $i \in I \cap I'$. Assumption (P2) reads that $p'_{ab} = p_{ab}$, for $a, b \in \{0, 1\}$, so we use the estimates $\hat{p}_{ab}$ of p_{ab} based on $I_{\text{test}} \subset I$. It follows that $\hat{\alpha}'_p = \left(\hat{P}^T\right)^{-1} \hat{\alpha}'_m$.

Prior probability shift can be quantified by the difference $\delta := \alpha' - \alpha$, which we will briefly refer to as the *drift*. The double sampling scheme (see Subsection 2.3) corresponds to the special case $\delta = 0$. In Section 3, we analytically derive expressions for the MSE of the misclassification estimator and calibration estimator when $\delta \neq 0$.

3. Analytical Results

In this section, we provide new analytical derivations of the bias and variance of (1) the misclassification estimator and (2) the calibration estimator. The resulting expressions for the bias and variance of the misclassification estimator are presented in Subsection 3.1. Those of the calibration estimator are presented in Subsection 3.2. Moreover, new sharp upper and lower bounds for the bias of calibration estimator are obtained.

The analytical derivations are rather long. Therefore, we have decided to include only a description of the proof strategy in the main text. The full details of the derivations can be found in the Appendix (Section 6).

3.1. Bias and Variance of the Misclassification Estimator

Expressions for the bias B and variance V of the misclassification estimator a_p under drift δ can be derived easily from the expressions presented by Kloos et al. (2020). It follows that

$$B\left[\hat{\alpha}'_p\right] = \frac{1}{n\left(p_{00} + p_{11} - 1\right)^2} \cdot \left[\frac{\alpha'}{\alpha} p_{11}\left(1 - p_{11}\right) - \frac{1 - \alpha'}{\alpha} p_{00}(1 - p_{00})\right] + O\left(\frac{1}{n^2}\right) \tag{6}$$

$$= \frac{p_{00} - p_{11}}{n\left(p_{00} + p_{11} - 1\right)} + \frac{\delta}{n\left(p_{00} + p_{11} - 1\right)^2} \cdot \left(\frac{p_{11}\left(1 - p_{11}\right)}{\alpha} + \frac{p_{00}\left(1 - p_{00}\right)}{1 - \alpha}\right) + O\left(\frac{1}{n^2}\right),$$

which is increasing in δ. We note that the bias might be negative when $\delta = 0$, so the absolute value of the bias might first decrease for increasing δ. The variance of the misclassification estimator equals

$$V\left(\hat{\alpha}'_p\right) = \frac{\left(1-\alpha'\right)^2 V\left(\hat{p}_{00}\right) + \alpha'^2 V\left(\hat{p}_{11}\right)}{\left(p_{00}+p_{11}-1\right)^2} + O\left(\frac{1}{n^2}\right). \tag{7}$$

We neglect the terms of order $1/n^2$ and use Expressions (14) and (15) from the Appendix to obtain

$$V\left(\hat{\alpha}'_p\right) = \frac{1}{n\left(p_{00}+p_{11}-1\right)^2} \cdot \Big[T + 2\delta(p_{00}-p_{11})(p_{00}+p_{11}-1) + \delta^2 \cdot \left(\frac{p_{11}(1-p_{11})}{\alpha} + \frac{p_{00}(1-p_{00})}{1-\alpha}\right)\Big] + O\left(\frac{1}{n^2}\right), \tag{8}$$

in which $T := (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$. If $p_{00} \geq p_{11}$, then the variance increases as the drift δ increases. If $p_{00} < p_{11}$, then the effect of the drift is not immediately clear: a larger value of δ might result in a lower variance, depending on the values of a and δ. In Section 4, we will numerically analyse the behaviour of $V(\hat{\alpha}'_p)$ as a function of α and δ.

3.2. Bias and Variance of the Calibration Estimator

Kloos et al. (2020) describe their proof strategy as follows. The expressions for the bias and variance of the calibration estimator are derived by conditioning on the base rate in the target population. However, if the drift δ is nonzero, the proof strategy chosen breaks down. Therefore, we have adapted the proof strategy to hold for nonzero δ, resulting in the following expressions (see Expressions (9) and (10)).

Theorem 1. *The bias of $\hat{\alpha}'_c$ as estimator of α under drift δ is given by*

$$B\left[\hat{a}'_c\right] = \delta \frac{T}{\beta(1-\beta)} + O\left(\frac{1}{n^2}\right), \tag{9}$$

in which $\beta := (1-\alpha)(1-p_{00}) + \alpha p_{11}$ and $\mathrm{T} = (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$. With that notation, the variance of $\hat{\alpha}'_c$, under drift δ, is given by

$$V(\hat{a}'_c) = \frac{\alpha(1-\alpha)}{n}\left[\frac{T}{\beta(1-\beta)} + 2\delta\left(p_{00}+p_{11}-1\right)\left(\frac{p_{11}(1-p_{00})}{\beta^2} - \frac{p_{00}(1-p_{11})}{(1-\beta)^2}\right) + \delta^2\left(p_{00}+p_{11}-1\right)^2\left(\frac{p_{11}(1-p_{00})}{\beta^3} + \frac{p_{00}(1-p_{11})}{(1-\beta)^3}\right)\right] + O\left(\frac{1}{n^2}\right). \tag{10}$$

Proof strategy. The calibration estimator at time t_2 is given by $\hat{\alpha}'_c = \hat{\alpha}'_m \hat{c}_{11} + (1-\hat{\alpha}'_m)\hat{c}_{10}$. We first show that (the square of) the calibration estimator $\hat{\alpha}_m$ (at time t_1) and (the square of) $\hat{c}_{ab}$ are approximately uncorrelated. We use the results derived by Kloos et al. (2020) for the expectation and variance of the product $\hat{\alpha}_m \hat{c}_{ab}$. We then compute the expectation and variance of $\hat{c}_{ab}$, which is a ratio of two random variables. Moreover, the ratio has the form $X/(X+Y)$, so the numerator and denominator are not

independent. To compute the expectation (and variance) of such a ratio, we first choose a suitable random variable to condition on, then we need to apply Taylor's theorem to approximate the expectation of a ratio, and finally we apply the law of total expectation.

Consequently, we assume that (the square of) $\hat{\alpha}'_m$ (at time t_2) and (the square of) $\hat{c}_{ab}$ are approximately uncorrelated as well. We then derive the expressions for the bias and variance of the calibration estimator $\hat{\alpha}'_c$ as estimator of the base rate α'. The full proof can be found in the Appendix.

In contrast to the misclassification estimator, the calibration estimator is unbiased if $\delta = 0$. However, the calibration estimator may be biased strongly if the drift δ is nonzero. Using Expression (9), we are able to derive analytically the following sharp upper bound and sharp lower bound for the absolute value of the bias in terms of the absolute value of the drift.

Theorem 2. *The absolute value of the bias of $\hat{\alpha}'_c$ as estimator of $\alpha' = \alpha + \delta$ is bounded from above by $|\delta|$. If $p_{00} \leq p$ and $p_{11} \leq p$ for some $1/2 \leq p \leq 1$, then the absolute value of the bias is at least $4p(1-p)|\delta|$.*

Proof strategy. Theorem 1 states that the quadratic approximation of $B\left[\hat{\alpha}'_c\right]$ (in p_{00} and p_{11}) is a linear function in δ. The slope of that linear function is decreasing in both p_{00} and p_{11}, which can be shown by computing the partial derivatives of that slope with respect to p_{00} and p_{11}. The statement of Theorem 2 follows immediately. The full proof can be found in the Appendix.

In summary, the bias of the misclassification estimator is of order $1/n$ while that of the calibration estimator does not decrease for increasing n. The implication is that the conclusions drawn by Kloos et al. (2020), assuming the double sampling scheme (see assumption (D1) in Subsection 2.3), do not hold when assuming nonzero prior probability shift (see assumptions (P1) and (P2) in Subsection 2.4). More specifically, if the drift δ is nonzero, our analytical results indicate that a decision boundary arises that depends on the size n of the test set. The aim of Section 4 is to investigate the properties of the decision boundary.

4. Numerical Results

The analytical results from Section 3 indicate that, in case δ is nonzero, a decision boundary arises (between preferring (1) the misclassification estimator and (2) the calibration estimator to reduce misclassification bias). The aim of this section is to understand that decision boundary and use it to provide a practical recommendation on machine learning applications in official statistics. The latter is the main focus of Subsection 4.3. In advance, we investigate the bias of the calibration estimator more closely in Subsection 4.1 and the difference in MSE between the two estimators in Subsection 4.2. To facilitate the use of our practical recommendation, we include a brief example in Subsection 4.4.

4.1. Bias of the Calibration Estimator

We start plotting $T/(\beta(1-\beta))$, the absolute value of the slope of the bias of the calibration estimator, as a function of the probability of correct classification for different values of α, that is, the base rate in the test set. For visualisation purposes, we restrict the function to

$p_{00} = p_{11}$, parameterised by p. The results are depicted in Figure 1, including the theoretical lower bound stated in Theorem 2. The slope of the bias as a function of p is decreasing from 1 at $p = 0.5$ to 0 at $p = 1$. The smaller the value of a, the later the function drops to 0. The reason is that the drift δ is defined as an absolute number and therefore it is relatively larger for smaller values of α. From this observation we may conclude that the impact of (an absolute) drift δ on the bias of $\hat{\alpha}'_c$ increases if α is further away from 0.5, that is, if the so-called *class imbalance* increases.

4.2. *Difference in Mean Squared Error*

Subsequently, we investigate the difference $D(\hat{a}'_p, \hat{a}'_c) := MSE(\hat{a}'_p) - MSE(\hat{a}'_c)$ between the MSE of the misclassification estimator and that of the calibration estimator. The value of $D(\hat{a}'_p, \hat{a}'_c)$ as a function of δ is depicted in Figure 2 for each possible combination of $\alpha \in \{0.05, 0.3\}$, $n \in \{50,1000\}$ and $p_{00}, p_{11} \in \{0.6, 0.7\}$. Note that the drift δ ranges from $-\alpha$ to $1 - \alpha$, because $\alpha' = \alpha + \delta$ must lie between 0 and 1. We report the following four observations. First, the difference is positive if $\delta = 0$ in any of the line plots, which corresponds to the main conclusion drawn by Kloos et al. (2020). Second, when n is sufficiently large (thin lines), the difference between the line plots are small. The reason is that the contribution of the variance terms is negligible compared to that of the squared bias of $\hat{\alpha}'_c$, which does not depend on n (see Theorem 1). Third, for highly imbalanced data sets combined with small test sets, that is, α close to 0 and n small (thick dash-dotted lines), the variance of $\hat{\alpha}'_p$ dominates if either p_{00} is close to 0.5 or p_{11} is close to 0.5. As a result, the calibration estimator has the lowest MSE, independent of the magnitude of the drift δ. Fourth, if the class distribution is relatively

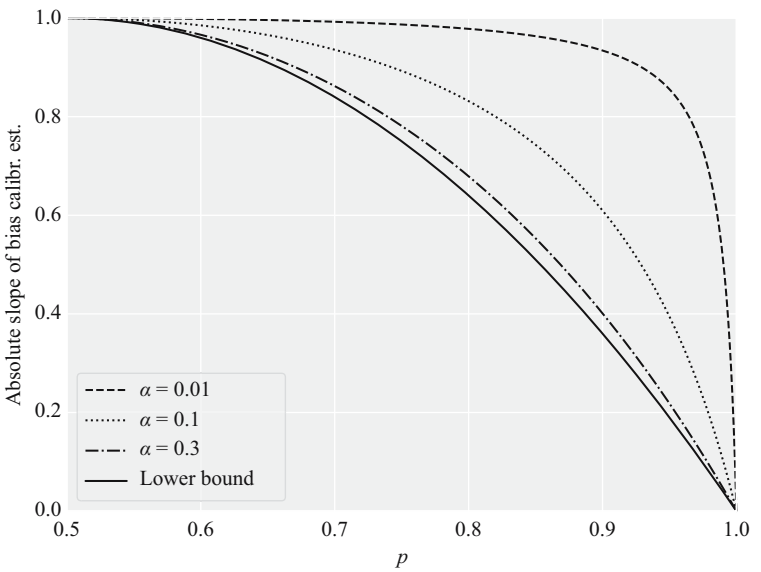

Fig. 1. The slope of the bias of the calibration estimator $\hat{\alpha}'_c$ as a function of the drift δ is equal to $-T/(\beta(1-\beta))$, which is strictly negative. The absolute value of that slope is plotted against the probability of correct classification p, assuming that $p_{00} = p_{11} = p$, for four different values of α. The solid black line depicts the theoretical lower bound (see Theorem 2) for the slope of the bias.

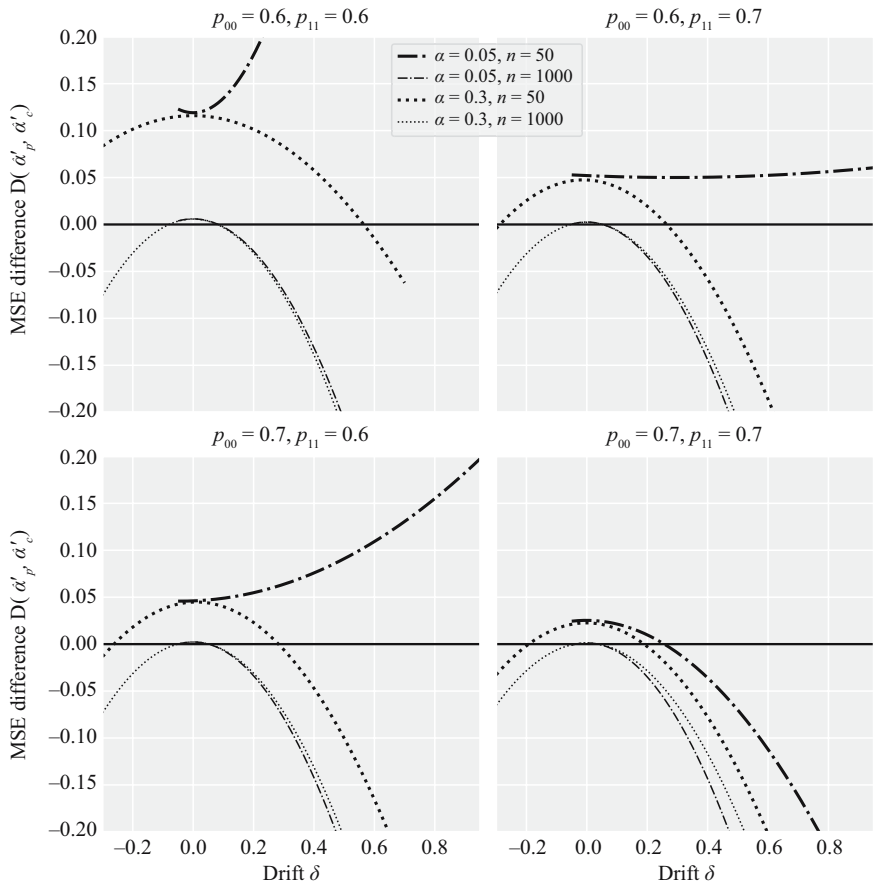

Fig. 2. The difference $D(\hat{\alpha}'_p, \hat{\alpha}'_c)$ between the MSE of the misclassification estimator $\hat{a}'_p$ and that of the calibration estimator $\hat{a}'_c$, plotted as a plotted as a function of the drift δ for each possible combination of $\alpha \in \{0.05, 0.3\}$, $n \in \{50, 1000\}$ and $p_{00}, p_{11} \in \{0.6, 0.7\}$. Note that the drift δ ranges from $-\alpha$ to $1-\alpha$, because $\alpha' = \alpha + \delta$ must lie between 0 and 1.

balanced (dotted lines), the difference $D\left(\hat{\alpha}'_p, \hat{\alpha}'_c\right)$ will become negative if δ increases, but the intersection moves farther away from $\delta = 0$ as n decreases.

4.3. The Decision Boundary

Here, we numerically compute the unique positive value δ^* (if it exists) at which the MSE of the misclassification and calibration estimator are identical. That is, we collect and reorganise the points of intersection $D(\hat{\alpha}'_p, \hat{\alpha}'_c) = 0$ as discussed in Subsection 4.2. We set $p_{00} = p_{11}$ and view $D(\hat{\alpha}'_p, \hat{\alpha}'_c)$ as a map from $\mathbb{R}^2$ to $\mathbb{R}$ by fixing α and n and using δ and $p := p_{00} = p_{11}$ as variables. Then, we define $\delta^*(p)$ as the positive solution $\delta > 0$ of the equation $D(\hat{\alpha}'_p, \hat{\alpha}'_c) = 0$ for p fixed, if the solution exists. Otherwise, we set $\delta^*(p) = 1$, which occurs when α, n, and p are all small. For an example, see the thick dash-dotted line ($\alpha = 0.05$ and $n = 50$) in the top left panel ($p_{00} = p_{11} = 0.6$) in Figure 2. The obtained function $p \mapsto \delta^*(p)$ is shown in Figure 3.

Interestingly, the result is a decreasing function of p. At first sight, the result might seem to contradict the result obtained in the first analysis, cf. Figure 1. There, the absolute slope of the bias as function of δ decreases with increasing p. Hence, the MSE of $\hat{\alpha}'_c$ increases more slowly as a function of δ with increasing p. However, the result in Figure 3 follows from the fact that the difference in variance between $\hat{\alpha}'_c$ and $\hat{\alpha}'_p$rapidly decreases as p increases.

We stress that the lines in Figure 3 can be interpreted as decision boundaries. Each statistical indicator that is based on a classification algorithm plots somewhere in the (p, δ)-plane depicted in Figure 3. If the plot of the indicator in the (p, δ)-plane ends up above the decision boundary (which depends on α and n), then the misclassification estimator should be preferred over the calibration estimator to reduce misclassification bias. Otherwise, the calibration estimator should be preferred. In Subsection 4.4 below, we describe this approach in more detail.

4.4. The Recommended Approach and an Illustrative Example

The analysis of the decision boundary, in Subsection 4.3 above, shows that the decision for an optimal correction method depends on the true value of the drift δ. In practice, the drift δ cannot be estimated before estimating α and α', but an official statistician will use his or her intuition about the possible size of δ. We therefore recommend the following approach.

First, estimate p_{00} and p_{11} using the test set of size n and estimate a (at time t_1) by $\hat{\alpha}_c$. Then, use the analytical results in Section 3 to compute δ^* as depicted in Figure 3. Finally, consider if δ could be larger than δ^*. If so, use the misclassification estimator $\hat{\alpha}'_p$ to estimate α'. Otherwise, use the calibration estimator $\hat{\alpha}'_c$ to estimate α'.

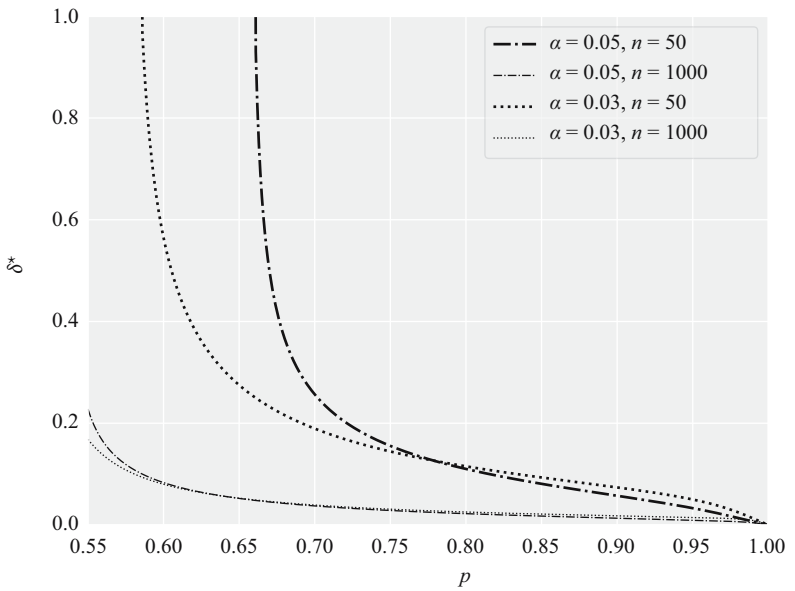

Fig. 3. The unique positive value δ^* (if it exists) for which $D(\hat{\alpha}'_p, \hat{\alpha}'_c) = 0$, as a function of the probability of correct classication p, assuming $p_{00} = p_{11} = p$. The lines should be interpreted as decision boundaries: below each of these lines the calibration estimator is preferred, while above each of the lines the misclassication estimator is preferred.

We illustrate the recommended approach by including the following example. Consider two populations I (population at time t_1) and I' (population at time t_2) that contain at least one million objects each. Draw a random sample $I_{\text{test}} \subset I$ of size $n = 500$ from the population at time t_1. Based on that random sample, referred to as the test set, we estimate the probabilities of correct classification. We assume that they result in $\hat{p}_{00} = 0.7$ and $\hat{p}_{11} = 0.8$. Following the recommendation by Kloos et al. (2020), we apply the calibration estimator $\hat{\alpha}_c$ to population I and obtain an estimate of α (at time t_1). We assume that the result equals $\hat{\alpha}_c = 0.2$.

Next, we estimate the MSE of $\hat{\alpha}'_p$ at time t_2 by plugging in the values of n, $\hat{\alpha}_c$, $\hat{p}_{00}$, and $\hat{p}_{11}$ into Expressions (6) and (8), yielding

$$\widehat{MSE}\left(\hat{\alpha}'_p\right) \approx \hat{V}\left(\hat{\alpha}'_p\right) \approx 0.0085 \cdot \delta^2 - 0.0008 \cdot \delta + 0.0016. \tag{11}$$

In the above computation, we have neglected terms of order $1/n^2$ and smaller. Similarly, we find

$$\widehat{MSE}\left(\hat{\alpha}'_c\right) \approx 0.694796 \cdot \delta^2 + 0.000356 \cdot \delta + 0.00267. \tag{12}$$

It follows that

$$\hat{D}\left(\hat{\alpha}'_p, \hat{\alpha}'_c\right) \approx -0.686296 \cdot \delta^2 - 0.001156 \cdot \delta + 0.001333. \tag{13}$$

Straightforward computation results in $\delta^* \approx 0.043$ (and the other root of (13) equals -0.045).

Thus, in the example above, our recommendation is to use the calibration estimator if δ can be assumed to lie between -0.045 and 0.043. Otherwise, we recommend to use the misclassification estimator. Although neither the true value nor an estimate of δ is known a priori, we believe that computing δ^* (and estimating $D(\hat{\alpha}'_p, \hat{\alpha}'_c)$ as function of δ) provides an official statistician with sufficient information to select one of the two bias correction methods in practice.

We conclude this section by two remarks. Our first remark is that one should always compute the (estimated) bias and variance of the applied estimator, for they might still be high, for example, when n and p are small. Our second remark is that the recommended approach should be adopted if the misclassification estimator and calibration estimator are the only two estimators under consideration. We admit that there may exist other estimators that might reduce misclassification bias even further.

5. Conclusions and Discussion

In this article, we investigated the statistical output quality of official statistics that are based on classification algorithms. The problem statement reads: how to reduce misclassification bias of classifier-based statistics that are affected by prior probability shift? In our research, we focused on two bias correction methods, namely (1) the misclassification estimator and (2) the calibration estimator. So far, the results known for these two estimators under the double sampling scheme (see Kloos et al. 2020) fail to hold

under prior probability shift. We have examined the MSE of the two estimators under prior probability shift, resulting in the following three contributions.

The first contribution of the article consists of expressions of the bias and variance of the two estimators when assuming prior probability shift (see Section 3). They extend the derivations by Kloos et al. (2020), for we obtain their results by setting the drift parameter δ to 0. Moreover, our analytical results show that (1) the calibration estimator is no longer unbiased and that (2) the quadratic approximation (in p_{00} and p_{11}) to the bias is a linearly decreasing function of the drift δ and does not depend on the test set size n. The second contribution of the article is that we present a decision boundary for choosing between the two estimators (see Section 4). The decision boundary depends on (a) the model accuracy measured by p_{11} and p_{00}, (b) the class distribution measured by α and (c) the test set size n. This specific result provides a better understanding of the output quality of official statistics based on machine learning algorithms. The third contribution of the article is our practical recommendation for choosing between the two estimators in practice (see Subsection 4.4).

The main conclusion of the article, enveloping the three contributions, is that if either (A) the performance of the classifier (in terms of p_{00} and p_{11}) is relatively low or (B) the drift δ is close enough to 0, then the calibration estimator should be preferred over the misclassification estimator. In earlier studies, this distinction has never been made. Moreover, our results show the impact of the size and frequency of the training and test data sets on output quality. Essentially, we show that an official statistician should be careful when applying the calibration estimator to time series data, unless training and test data in each publication period are available to retrain the classifier and adapt it to concept drift.

In the event that concept drift adaptation is considered too expensive, given particular cost constraints, the main conclusion (see above) implies that some minimal classification accuracy is required in order to use the misclassification estimator. In general, more labelled training data have to be created to guarantee higher classification accuracy. In other words, NSIs should be vigilant when evaluating the cost efficiency of implementing machine learning algorithms in the production process of official statistics. In the end, a substantial amount of high quality annotated data have to be created manually and consistently over a long period of time, which requires long-term investments in data analysts and domain experts.

Finally, our results suggest four directions for future research. First, the robustness of classifier-based estimators should also be investigated for other types of concept drift, starting with the less restrictive type of prior probability shift as defined by Webb et al. (2016). Second, it might be worthwhile to examine methods for concept drift adaptation that are based partly on unlabelled data, by carefully incorporating changes in the distribution of $P(X)$. Third, combinations or ensembles of different estimators require further research. We believe that a well-chosen combination of estimators will increase the overall robustness of classifier-based estimators under concept drift. Fourth, we recommend to investigate applications of machine learning in official statistics that are more complicated than applications involving the base rate. An ambitious goal is to understand the effect of misclassification bias and concept drift in applications of machine learning to panel-based surveys, such as estimating panel attrition (Liu 2020) or nonresponse adjustments (Buskirk and Kolenikov 2015).

6. Appendix

This appendix contains the proofs of the theorems presented in this article. First, we remark that Kloos et al. (2020) have shown that $\hat{p}_{00}$ and $\hat{p}_{11}$ are uncorrelated, and that the variance of $\hat{p}_{11}$ is equal to

$$V(\hat{p}_{11}) = \frac{p_{11}(1-p_{11})}{n\alpha}\left[1+\frac{1-\alpha}{n\alpha}\right] + O\left(\frac{1}{n^3}\right). \tag{14}$$

Similarly, the variance of $\hat{p}_{00}$ is given by

$$V(\hat{p}_{00}) = \frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1+\frac{\alpha}{n(1-\alpha)}\right] + O\left(\frac{1}{n^3}\right). \tag{15}$$

The proof of Theorem 1 relies on the following lemma.

Lemma 1. *The expectation and variance of $\hat{c}_{11}$ are given by*

$$\mathbb{E}[\hat{c}_{11}] = \frac{\alpha p_{11}}{\beta} \text{ and } V(\hat{c}_{11}) = \frac{\alpha(1-\alpha)}{n}\frac{p_{11}(1-p_{00})}{\beta^3}, \tag{16}$$

in which $\beta = (1-\alpha)(1-p_{00}) + \alpha p_{11}$ as before. Similarly, the expectation and variance of $\hat{c}_{10}$ are given by

$$\mathbb{E}[\hat{c}_{10}] = \frac{\alpha(1-p_{10})}{1-\beta} \text{ and } V(\hat{c}_{10}) = \frac{\alpha(1-\alpha)}{n}\frac{p_{00}(1-p_{11})}{(1-\beta)^3}. \tag{17}$$

Moreover, the covariance of $\hat{c}_{11}$ and $\hat{c}_{10}$ satisfies $C(\hat{c}_{11}, \hat{c}_{10}) = O(1/n^2)$.

Proof of Lemma 1. We will first compute the expectation and variance of $\hat{c}_{11}$. The derivations of the expectation and variance of $\hat{c}_{10}$ are similar to those of $\hat{c}_{11}$ and are therefore not included. At the end, we show that the covariance of $\hat{c}_{11}$ and $\hat{c}_{10}$ is equal to 0.

Expectation. To compute the expectation of $\hat{c}_{11}$, we condition on $n_{1+} := n_{11} + n_{10}$. Note that $n_{0+} = n - n_{1+}$ is known as soon as n_{1+} is known. It holds that $\hat{c}_{11} \mid n_{1+} \overset{d}{=} \mathrm{X}/(X+\mathrm{Y})$, with $X \sim Bin(n_1 + , p_{11})$ and $Y \sim Bin(n_0 + , 1 - p_{00})$. We introduce the random variable $\beta_+ := n_{1+}p_{11} + n_{0+}(1-p_{00})$. A second-order Taylor approximation then yields

$$\mathbb{E}[\hat{c}_{11}|n_{1+}] = \frac{n_1 + p_{11}}{\beta_+} - \frac{n_{0+}(1-p_{00})}{\beta_+^3}n_1 + p_{11}(1-p_{11}) + \frac{n_1+p_{11}}{\beta_+^3}n_0 + p_{00}(1-p_{00})$$

$$+ O\left(\frac{1}{n^2}\right) = \frac{n_1+p_{11}}{\beta_+} + p_{11}(1-p_{00})(p_{00}+p_{11}-1)\frac{n_0+n_{1+}}{\beta_+^3} + O\left(\frac{1}{n^2}\right). \tag{18}$$

We then introduce the random variable $Z \sim Bin(n, \alpha)$ (that is, $Z \stackrel{d}{=} n_{1+}$). Applying a Taylor approximation to the first term of Expression (18) yields

$$\begin{aligned}\mathbb{E}\left[\frac{n_1 + p_{11}}{\beta_+}\right] &= \mathbb{E}\left[\frac{p_{11}Z}{n(1-p_{00}) + (p_{00}+p_{11}-1)Z}\right] \\ &= \frac{\alpha p_{11}}{\beta} - \frac{1}{2}\frac{2np_{11}(1-p_{00})(p_{00}+p_{11}-1)}{n^3\beta^3}n\alpha(1-\alpha) + O\left(\frac{1}{n^2}\right) \\ &= \frac{\alpha p_{11}}{\beta} - \frac{\alpha(1-\alpha)}{n}\frac{p_{11}(1-p_{00})(p_{00}+p_{11}-1)}{\beta^3} + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{19}$$

Next, apply a Taylor approximation to (the stochastic part of) the second term in (18):

$$\mathbb{E}\left[\frac{Z(n-Z)}{\beta_+^3}\right] = \frac{\alpha(1-\alpha)}{n\beta^3} + O\left(\frac{1}{n^2}\right). \tag{20}$$

The law of total expectation, combining (19) and (20), results in

$$\mathbb{E}[\hat{c}_{11}] = \frac{\alpha p_{11}}{\beta} + O\left(\frac{1}{n^2}\right). \tag{21}$$

Similarly, it can be shown that

$$\mathbb{E}[\hat{c}_{10}] = \frac{\alpha(1-p_{11})}{1-\beta} + O\left(\frac{1}{n^2}\right), \tag{22}$$

Variance. We compute $V(\hat{c}_{11})$ as $\mathbb{E}\left[\hat{c}_{11}^2\right] - \mathbb{E}[\hat{c}_{11}]^2$, because we have already derived an expression for the latter term. The random variable $\hat{c}_{11}^2 | n_{1+}$ is distributed as $X^2/(X+Y)^2$, with X and Y as before. Setting $f(x,y) = x^2/(x+y)^2$ yields the partial derivatives (twice in x and twice in y)

$$f_{xx}(x,y) = \frac{2y^2 - 4xy}{(x+y)^4}, \quad \textit{and} \quad f_{yy}(x,y) = \frac{6x^2}{(x+y)^4}. \tag{23}$$

It follows, neglecting terms of higher order, that

$$\begin{aligned}\mathbb{E}\left[\hat{c}_{11}^2 | n_{1+}\right] &\approx \frac{n_1^2 + p_{11}^2}{\beta_+^2} + \frac{n_{0+}^2(1-p_{00})^2 - 2n_{1+}n_{0+}p_{11}(1-p_{00})}{\beta_+^4} \\ n_{1+}p_{11}(1-p_{11}) + \frac{3n_1^2 + p_{11}^2}{\beta_+^4}n_{0+}p_{00}(1-p_{00}) &= \frac{n_1^2 + p_{11}^2}{\beta_+^2} + p_{11}(1-p_{00}) \\ \frac{n_{1+}n_{0+}\left(n(1-p_{00})(1-p11) + n_{1+}(p_{00}+p_{11}-1)(2p_{11}+1)\right)}{\beta_+^4}&. \end{aligned} \tag{24}$$

Again, let $Z \sim Bin(n, \alpha)$ and consider the function $f(z) = z^2/(A + Bz)^2$, with $A = n(1-p_{00})$ and $B = (p_{00} + p_{11} - 1)$. The partial derivative twice in z equals

$$f_{zz}(z) = \frac{2A^2 - 4ABz}{(A+Bz)^4}. \tag{25}$$

The conditional expectation then equals (up to terms of order $1/n^2$)

$$\mathbb{E}\left[\frac{n_{1+}^2 p_{11}^2}{\beta_+^2}\right] = \mathbb{E}\left[\frac{p_{11}^2 Z^2}{(A+BZ)^2}\right] = \frac{\alpha^2 p_{11}^2}{\beta^2}$$

$$+ p_{11}^2 \frac{n^2(1-p_{00})^2 - 2n^2\alpha\big(1-p_{00}\big)\big(p_{00}+p_{11}-1\big)}{n^4\beta^4} n\alpha(1-\alpha) + O\left(\frac{1}{n^2}\right)$$

$$= c_{11}^2 + \frac{\alpha(1-\alpha)}{n}\frac{p_{11}^2\big(1-p_{00}\big)\big(1-p_{00}-2\alpha\big(p_{00}+p_{11}-1\big)\big)}{\beta^4} + O\left(\frac{1}{n^2}\right). \tag{26}$$

Apply a Taylor approximation to (the stochastic part of) the second term in (24) to obtain

$$\frac{\alpha(1-\alpha)}{n}\frac{p_{11}\big(1-p_{00}\big)\big((1-p_{00})(1-p_{11})+\alpha\big(p_{00}+p_{11}-1\big)(2p_{11}+1)\big)}{\beta^4} + O\left(\frac{1}{n^2}\right). \tag{27}$$

At last, combining (26) and (27), and subtracting (21) squared, the variance of $\hat{c}_{11}$ can be expressed as

$$V(\hat{c}_{11}) = \frac{\alpha(1-\alpha)}{n}\frac{p_{11}(1-p_{00})}{\beta^3} + O\left(\frac{1}{n^2}\right). \tag{28}$$

Similarly, it can be shown that

$$V(\hat{c}_{10}) = \frac{\alpha(1-\alpha)}{n}\frac{p_{00}(1-p_{11})}{(1-\beta)^3} + O\left(\frac{1}{n^2}\right). \tag{29}$$

Covariance. We compute the covariance of $\hat{c}_{11}$ and $\hat{c}_{10}$ as $C(\hat{c}_{11},\hat{c}_{10}) = \mathbb{E}[\hat{c}_{11}\hat{c}_{10}]$-$\mathbb{E}[\hat{c}_{11}]$ $\mathbb{E}[\hat{c}_{10}]$. It remains to compute $\mathbb{E}[\hat{c}_{11}\hat{c}_{10}]$. The strategy is similar to that of the computation of $\mathbb{E}\big[\hat{c}_{11}^2\big]$, so the details are not included. In brief, we again condition on n_{1+} and note that $\hat{c}_{11}\hat{c}_{10}|n_{1+}$ is distributed as $X(n_{1+}-X)/((X+Y)(n-(X+Y))$, with X and Y as before. Apply a second-order Taylor approximation to compute the conditional expectation $\mathbb{E}[\hat{c}_{11}\hat{c}_{10}|\ n_{1+}]$. Then, use that n_{1+} is distributed as $Z \sim Bin(n, \alpha)$ and apply the law of total expectation to find

$$\mathbb{E}[\hat{c}_{11}\hat{c}_{10}] = \frac{\alpha(1-\alpha)p_{11}(1-p_{11})}{\beta(1-\beta)} + O\left(\frac{1}{n^2}\right). \tag{30}$$

Combined with expressions (21) and (21) this concludes the proof of $C(\hat{c}_{11},\hat{c}_{10}) = O(1/n^2)$.

We will now provide the proof of Theorem 1 below.

Proof of Theorem 1. We will compute the bias and variance of the calibration estimator $\hat{\alpha}_c'$ as estimator of the base rate α' at time t_2.

Bias. Recall that the calibration estimator $\hat{\alpha}_c$ at time t_1 is given by

$$\hat{\alpha}_c = \hat{\alpha}_m\hat{c}_{11} + (1-\hat{\alpha}_m)\hat{c}_{10}. \tag{31}$$

The expectation of the product $\hat{\alpha}_m \hat{c}_{11}$ has been derived by Kloos et al. (2020) and equals

$$\mathbb{E}[\hat{\alpha}_m \hat{c}_{11}] = \alpha p_{11}. \tag{32}$$

The expectation of $\hat{\alpha}_m$ equals $\mathbb{E}[\hat{\alpha}_m] = (1-\alpha)(1-p_{00}) + \alpha p_{11} =: \beta$. Lemma 1 then implies

$$\mathbb{E}[\hat{\alpha}_m \hat{c}_{11}] = \mathbb{E}[\hat{\alpha}_m]\mathbb{E}[\hat{c}_{11}] + O\left(\frac{1}{n^2}\right). \tag{33}$$

Hence, $\hat{\alpha}_m$ and $\hat{c}_{11}$ are approximately uncorrelated. Similarly, $\hat{\alpha}_m$ and $\hat{c}_{10}$ are approximately uncorrelated as well.

Therefore, we assume that $\hat{\alpha}'_m$ (at time t_2) and $\hat{c}_{ab}$ (at time t_1), for $a, b \in \{0, 1\}$, are also approximately uncorrelated. Hence,

$$\mathbb{E}\left[\hat{\alpha}'_c\right] = \mathbb{E}\left[\hat{\alpha}'_m\right]\mathbb{E}[\hat{c}_{11}] + \mathbb{E}\left[1 - \hat{\alpha}'_m\right]\mathbb{E}[\hat{c}_{10}]. \tag{34}$$

We introduce the notation $\beta' := (1-\alpha')(1-\mathrm{p}_{00}) + \alpha' p_{11} = \mathbb{E}\left[\hat{\alpha}'_m\right]$. Substituting $\alpha' = \alpha + \delta$ and neglecting terms of order $1/n^2$ yields

$$\begin{aligned}
\mathbb{E}\left[\hat{\alpha}'_c\right] &= \beta' \frac{\alpha p_{11}}{\beta} + (1-\beta')\frac{\alpha(1-p_{11})}{1-\beta} \\
&= \alpha p_{11} + \delta(p_{00} + p_{11} - 1)\frac{\alpha p_{11}}{\beta} + \alpha(1-p_{11}) + \delta(1 - p_{00} - p_{11})\frac{\alpha(1-p_{11})}{1-\beta} \\
&= \alpha + \frac{\delta\alpha}{\beta(1-\beta)}\big((1-\beta)p_{11} - \beta(1-p_{11})\big)(p_{00} + p_{11} - 1) \\
&= \alpha + \frac{\delta\alpha(1-\alpha)(p_{00} + p_{11} - 1)^2}{\beta(1-\beta)}.
\end{aligned} \tag{35}$$

It is straightforward to check that

$$\beta(1-\beta) - \alpha(1-\alpha)(p_{00} + p_{11} - 1)^2 = \alpha p_{11}(1-p_{11}) + (1-\alpha)p_{00}(1-p_{00}) =: T. \tag{36}$$

Hence,

$$\mathbb{E}\left[\hat{\alpha}'_c\right] = \alpha + \delta\left(\frac{\beta(1-\beta) - T}{\beta(1-\beta)}\right) + O\left(\frac{1}{n^2}\right) = \alpha' - \delta\frac{T}{\beta(1-\beta)} + O\left(\frac{1}{n^2}\right). \tag{37}$$

Thus, we may conclude that the bias of $\hat{\alpha}'_c$ as estimator of α' is equal to

$$B[\hat{\alpha}'_c] = \delta\frac{T}{\beta(1-\beta)} + O\left(\frac{1}{n^2}\right). \tag{38}$$

Variance. The variance of the product $\hat{\alpha}_m \hat{c}_{11}$ has been derived by Kloos et al. (2020) and equals

$$V(\hat{a}_m \hat{c}_{11}) = \frac{\beta}{n}\frac{\alpha p_{11}}{\beta}\left(1 - \frac{\alpha p_{11}}{\beta}\right) + O\left(\frac{1}{n^2}\right) = \frac{\alpha(1-\alpha)}{n}\frac{p_{11}(1-p_{00})}{\beta} + O\left(\frac{1}{n^2}\right). \tag{39}$$

Lemma 1 then implies that

$$V(\hat{a}_m\hat{c}_{11}) = (\mathbb{E}[\hat{\alpha}_m])^2 V(\hat{c}_{11}) + O\left(\frac{1}{n^2}\right). \tag{40}$$

Moreover, note that

$$\mathbb{E}[(\hat{\alpha}'_m)^2] = \mathbb{E}[\hat{\alpha}'_m]^2 + V(\hat{\alpha}'_m) = \mathbb{E}[\hat{\alpha}'_m]^2 + O\left(\frac{1}{N}\right). \tag{41}$$

The above (combined with $n \lll N$) proves that

$$\mathbb{E}[(\hat{\alpha}'_m\hat{c}_{11})^2] - \mathbb{E}[(\hat{\alpha}'_m)^2]\mathbb{E}[(\hat{c}_{11})^2] = O\left(\frac{1}{n^2}\right). \tag{42}$$

Thus, we conclude that (1) $\hat{\alpha}_m^2$ and $\hat{c}_{11}^2$ are approximately uncorrelated. Similarly, it can be shown that (2) $\hat{\alpha}_m^2$ and $\hat{c}_{10}^2$ are approximately uncorrelated and that (3) $\hat{\alpha}_m(1-\hat{\alpha}_m)$ and $\hat{c}_{11}\hat{c}_{10}$ are approximately uncorrelated.

Therefore, we assume that (1)-(3) still hold when $\hat{\alpha}_m$ is replaced by $\hat{\alpha}'_m$. These three assumptions, combined with Expression (41), imply that

$$\begin{aligned} V(\hat{\alpha}'_c) &= V(\hat{\alpha}'_m\hat{c}_{11}) + V\big((1-\hat{\alpha}'_m)\hat{c}_{10}\big) + C(\hat{\alpha}'_m\hat{c}_{11}, (1-\hat{\alpha}'_m)\hat{c}_{10}) \\ &= \mathbb{E}\big[\hat{\alpha}'_m\big]^2 V(\hat{c}_{11}) + \mathbb{E}\big[1-\hat{\alpha}'_m\big]^2 V(\hat{c}_{10}) + \mathbb{E}\big[\hat{\alpha}'_m\big]\mathbb{E}\big[\big(1-\hat{\alpha}'_m\big)\big] C(\hat{c}_{11}, \hat{c}_{10}), \end{aligned} \tag{43}$$

where we have neglected terms of order $1/N$ and $1/n^2$. We substitute $\mathbb{E}\big[\hat{\alpha}'_m\big] = \beta'$ in the above and use the expressions for $V(\hat{c}_{11})$, $V(\hat{c}_{10})$ and $C(\hat{c}_{11}, \hat{c}_{10})$ from Lemma 1. We may conclude that

$$V(\hat{\alpha}'_c) = \frac{\alpha(1-\alpha)}{n}\left(\beta'^2\frac{p_{11}\big(1-p_{00}\big)}{\beta^3} + (1-\beta')^2\frac{p_{00}(1-p_{11})}{(1-\beta)^3}\right) + O\left(\frac{1}{n^2}\right). \tag{44}$$

Substituting $\alpha' = \alpha + \delta$ yields

$$\begin{aligned} V(\hat{\alpha}'_c) = \frac{\alpha(1-\alpha)}{n}\Bigg[&\frac{T}{\beta(1-\beta)} + 2\delta(p_{00}+p_{11}-1)\left(\frac{p_{11}(1-p_{00})}{\beta^2} - \frac{p_{00}(1-p_{11})}{(1-\beta)^2}\right) \\ &+ \delta^2(p_{00}+p_{11}-1)^2\left(\frac{p_{11}(1-p_{00})}{\beta^3} + \frac{p_{00}(1-p_{11})}{(1-\beta)^3}\right)\Bigg] + O\left(\frac{1}{n^2}\right). \end{aligned} \tag{45}$$

The expression above completes the derivation of the variance of the calibration estimator under prior probability shift. This concludes the proof of Theorem 1.

To prove the Theorem 2, we need the following lemma.

Lemma 2. *The slope of the absolute value of the quadratic approximation (in p_{00} and p_{11}) to the bias of the calibration estimator as a function of the absolute value $|\delta|$ of the prior probability shift is decreasing in p_{00} and p_{11} for all $1/2 \le p_{00} \le 1$ and $1/2 \le p_{11} \le 1$.*

Proof. We introduce the notation $x = p_{00},\ y = p_{11}$ and $\beta = \beta(x, y, \alpha) = (1-\alpha)(1-x) + \alpha y$. We then define the functions

$$f(x, y, \alpha) = \frac{(1-x)y}{\beta} \text{ and } g(x, y, \alpha) = \frac{x(1-y)}{1-\beta}. \tag{46}$$

The function $h = f + g$ then satisfies $|\delta| \cdot h(p_{00}, p_{11}, \alpha) = |B[\hat{\alpha}'_c]|$ up to terms of order $1/n^2$. We will examine the sign of the partial derivatives of h with respect to x and y, which we denote by h_x and h_y, respectively. To that end, we compute the partial derivatives of f and g, giving

$$f_x(x, y, \alpha) = \frac{-\alpha y^2}{\beta^2} \text{ and } g_x(x, y, \alpha) = \frac{\alpha(1-y)^2}{(1-\beta)^2}. \tag{47}$$

Hence, we obtain

$$h_x(x, y, \alpha) = \frac{\alpha}{\beta^2(1-\beta)^2} \cdot \Big(\big((1-y)\beta\big)^2 - \big(y(1-\beta)\big)^2\Big). \tag{48}$$

Setting this to zero yields $(1-y)^*\beta = y(1-\beta)$ or $(1-y)\beta = -y(1-\beta)$. As $1/2 \le x$, $y \le 1$ and $0 < \alpha < 1$ it follows that $1 - x \le \beta \le y$ with equality if and only if $1 - x = y$, that is, $x = y = 1/2$. It implies that $(1-y)\beta$ is nonnegative and that $y(1-\beta)$ is strictly positive, hence the Equation $(1-y)\beta = -y(1-\beta)$ has no solution. Moreover, it implies that $(1-y)\beta \le y(1-\beta)$ with equality only at $x = y = 1/2$. From this we may conclude that h is decreasing in x for all $1/2 < x \le 1$ and that $h_x(\frac{1}{2}, \cdot, \cdot) = 0$.

The partial derivatives h_x and h_y can be related through a simple symmetry argument: it holds that $\beta(y, x, \alpha) = 1 - \beta(x, y, 1-\alpha)$, which implies that $h(y, x, \alpha) = h(x, y, 1-\alpha)$. Consequently, it holds that $h_y(\cdot, \cdot, \alpha) = hx(\cdot, \cdot, 1-\alpha)$. It follows that h is also decreasing in y for all $1/2 < y \le 1$ and that $hy(\cdot, \frac{1}{2}, \cdot) = 0$.

We conclude that the slope h of the quadratic approximation (in p_{00} and p_{11}) to the bias of the calibration estimator under prior probability shift is decreasing in p_{00} and p_{11} for $1/2 < p_{00}, p_{11} < 1$, attaining its global maximum at $p_{00} = p_{11} = 1/2$, where $h = 1$ and $|B[\hat{\alpha}'_c]| = |\delta|$.

The statement of Theorem 2 is an immediate consequence of the lemma above.

Proof of Theorem 2. Lemma 2 implies that $|B[\hat{\alpha}'_c]| \le |\delta|$ and that $|B[\hat{\alpha}'_c]| \ge |\delta| \cdot h(p, p, \alpha)$. To simplify the latter, observe that $T(p, p, \alpha) = p(1-p)$ and that $0 \le 1 - p < \beta(p, p, \alpha) < p \le 1$, using that $1/2 \le p \le 1$ and $0 < \alpha < 1$. It follows that $\beta(1-\beta) \le 1/4$, which completes the proof.

7. References

Beck, M., F. Dumpert, and J. Feuerhake. 2018. *Machine learning in official statistics.* arXiv:1812.10422. DOI: https://doi.org/10.48550/arXiv.1812.10422.

Braaksma, B., and C. Zeelenberg. 2015. "Re-make/Re-model: Should big data change the modelling paradigm in official statistics?" *Statistical Journal of the IAOS* 31(2): 193–202. DOI: htpps://doi.org/10.3233/sji-150892.

Breiman, L. 2001. "Statistical modeling: The two cultures." *Statistical Science* 16(3): 199–231. DOI: htpps://doi.org/10.1214/ss/1009213726.

Bross, I.D.J. 1954. "Misclassification in 2 × 2 tables." *Biometrics* 10(4): 478–486. DOI: htpps://doi.org/10.2307/3001619.

Buelens, B., P.-P. de Wolf, and C. Zeelenberg. 2016. "Model based estimation at Statistics Netherlands." In European Conference on Quality in Official Statistics, Madrid, Spain. Available at: https://www.ine.es/q2016/docs/q2016Final00196.pdf.

Buonaccorsi, J.P. 2010. *Measurement Error: Models, Methods, and Applications.* Chapman & Hall/CRC, 31 May – 3 June, Boca Raton, Florida.

Buskirk, T.D., and S. Kolenikov. 2015. *Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification.* Available at: https://surveyinsights.org/?p = 5108 (accessed April 2020).

Costa, H, D. Almeida, F. Vala, F. Marcelino, and M. Caetano. 2018. "Land cover mapping from remotely sensed and auxiliary data for harmonized official statistics." ISPRS International Journal of Geo-Information 7(4):157. DOI: htpps://doi.org/10.3390/ijgi7040157.

Curier, R.L., T.J.A. de Jong, K. Strauch, K. Cramer, N. Rosenski, C. Schartner, M. Debusschere, H. Ziemons, D. Iren, and S. Bromuri. 2018. *Monitoring spatial sustainable development: Semi-automated analysis of satellite and aerial images for energy transition and sustainability indicators*. arXiv:1810.04881. DOI: https://doi.org/10.48550/arXiv.1810.04881.

Daas P.J.H., and S. van der Doef. 2020. "Detecting innovative companies via their website." *Statistical Journal of the IAOS* 36(4): 1239–1251. DOI: htpps://doi.org/10.3233/SJI-200627.

De Broe, S.M.M.G., P. Struijs, P.J.H. Daas, A. van Delden, J. Burger, J.A. van den Brakel, K.O. ten Bosch, C. Zeelenberg, and W.F.H. Ypma. 2020. *Updating the paradigm of official statistics.* CBDS Working Paper 02-20, Statistics Netherlands, The Hague/Heerlen.

European Commission. 2009. *Regulation of European Statistics*. Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri = CELEX%3A32009R0223 (accessed April 2020).

Eurostat. 2017. *European Statistics Code of Practice*. Available at: https://ec.europa.eu/eurostat/web/ (accessed April 2020).

Forman, G. 2015. "Counting positives accurately despite inaccurate classification." In *Machine Learning: ECML 2005, Lecture Notes in Computer Science*, edited by J. Gama, R. Camacho, P.B. Brazdil, A.M. Jorge, and L. Torgo: 564–575, Berlin, Heidelberg, Springer. DOI: https:/oi.org/10.1007/11564096_55.

Gama, J., I. Žliobaité, A. Bifet, M. Pechenizkiy, and A. Bouchachia. 2014. "A survey on concept drift adaptation." *ACM Computing Surveys* 46(4): 1–37. DOI: htpps://doi.org/10.1145/2523813.

Goldenberg, I., and G.I. Webb. 2019. "Survey of distance measures for quantifying concept drift and shift in numeric data." *Knowledge and Information Systems* 60(2): 591–615. DOI: https://doi.org/10.1007/s10115-018-1257-z.

González, P., A. Castaño, N.V. Chawla, and J.J. Del Coz. 2017. "A review on quantification learning." *ACM Computing Surveys* 50(5): 74:1–74:40. DOI: https://doi.org/10.1145/3117807.

Helmbold D.P., and P.M. Long. 1994. "Tracking drifting concepts by minimizing disagreements." *Machine Learning* 14(1): 27–45. DOI: https://doi.org/10.1007/ BF00993161.

Kenett, R.S., and G. Shmueli. 2016. "From quality to information quality in official statistics." *Journal of Official Statistics* 32(4): 867–885. DOI: https://doi.org/10.1515/-jos-2016-0045.

Kloos, K., Q.A. Meertens, S. Scholtus, and J.D. Karch. 2020. "Comparing correction methods to reduce misclassification bias." In *BNAIC/BENELEARN* 2020 edited by L. Cao, W.A. Kosters, and J. Lijffijt: 103–129, Leiden.

Kuha, J., and C.J. Skinner. 1997. "Categorical data analysis and misclassification." In *Survey Measurement and Process Quality*, edited by L.E. Lyberg, P.P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin: 633–670. Wiley, New York. DOI: https://doi.org/10.1002/9781118490013.

Liu, M. 2020. "Using machine learning models to predict attrition in a survey panel." In *Big Data Meets Survey Science*, edited by C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, and L.E. Lyberg: 415–433. John Wiley & Sons. doi: https://doi.org\10.1002/9781118976357.ch14.

Lu, J., A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. 2019. "Learning under concept drift: A review." *IEEE Transactions on Knowledge and Data Engineering* 31(12): 2346–2363. DOI: https://doi.org/10.1109/TKDE.2018.2876857.

Moreno-Torres, J.G., T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, and F. Herrera. 2012. "A unifying view on dataset shift in classification." *Pattern Recognition* 45(1): 521–530. DOI: https://doi.org/10.1016/j.patcog.2011.06.019.

O'Connor, B., R. Balasubramanyan, B.R. Routledge, and N.A. Smith. 2010. "From tweets to polls: Linking text sentiment to public opinion time series." In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)* May 23 – May 26, edited by M.A. Hearst: 122–129, Washington, D.C, U.S.A. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1536/1842.

OECD. 2011. *Quality Framework for OECD Statistical Activities*. Available at: https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm (accessed April 2020).

Schlimmer, J.C., and R.H. Granger. 1986. "Incremental learning from noisy data." *Machine Learning* 1(3): 317–354. DOI: https://doi.org/10.1007/BF00116895.

Scholtus, S., and A. van Delden. 2020. *The accuracy of estimators based on a binary classifier*. Discussion Paper 202006, Statistics Netherlands, The Hague. Available at: https://www.cbs.nl/-/media/_pdf/2020/06/classification-errors-binary.pdf.

Schwartz, J.E. 1985. "The neglected problem of measurement error in categorical data." *Sociological Methods & Research* 13(4): 435–466. DOI: https://doi.org/10.1177/0049124185013004001.

Tenenbein, A. 1970. "A double sampling scheme for estimating from binomial data with misclassifications." *Journal of the American Statistical Association* 65(331): 1350–1361. DOI: https://doi.org/10.1080/01621459.1970.10481170.

Van Delden, A., S. Scholtus, and J. Burger. 2016. "Accuracy of Mixed-Source Statistics as Affected by Classification Errors." *Journal of Official Statistics* 32(3): 619–642. DOI: https://doi.org/10.1515/jos-2016-0032.

Webb, G.I., R. Hyde, H. Cao, H.L. Nguyen, and F. Petitjean. 2016. "Characterizing concept drift." *Data Mining and Knowledge Discovery* 30(4): 964–994. DOI: https://doi.org/10.1007/s10618-015-0448-4.

Widmer, G., and M. Kubat. 1996. "Learning in the presence of concept drift and hidden contexts." *Machine Learning* 23(1): 69–101. DOI: https://doi.org/10.1023/A:1018046501280.

Received December 2020
Revised April 2021
Accepted June 2021