Comparing Correction Methods to Reduce Misclassification Bias

Kevin Kloos[1,5] [0000−0001−6980−4259], Quinten Meertens[3,4,5] [0000−0002−3485−8895], Sander Scholtus[5] [0000−0002−8316−8938], and Julian Karch[2] [0000−0002−1625−2822]

[1] Mathematical Institute, Leiden University, the Netherlands
[2] Institute of Psychology, Leiden University, the Netherlands
[3] Leiden Centre of Data Science, Leiden University, the Netherlands
[4] Center for Nonlinear Dynamics in Economics and Finance, University of Amsterdam, the Netherlands
[5] Statistics Netherlands, The Hague, the Netherlands †

Abstract. When applying supervised machine learning algorithms to classification, the classical goal is to reconstruct the true labels as accurately as possible. However, if the predictions of an accurate algorithm are aggregated, for example by counting the predictions of a single class label, the result is often still statistically biased. Implementing machine learning algorithms in the context of official statistics is therefore impeded. The statistical bias that occurs when aggregating the predictions of a machine learning algorithm is referred to as misclassification bias. In this paper, we focus on reducing the misclassification bias of binary classification algorithms by employing five existing estimation techniques, or estimators. As reducing bias might increase variance, the estimators are evaluated by their mean squared error (MSE). For three of the estimators, we are the first to derive an expression for the MSE in finite samples, complementing the existing asymptotic results in the literature. The expressions are then used to compute decision boundaries numerically, indicating under which conditions each of the estimators is optimal, i.e., has the lowest MSE. Our main conclusion is that the calibration estimator performs best in most applications. Moreover, the calibration estimator is unbiased and it significantly reduces the MSE compared to that of the uncorrected aggregated predictions, supporting the use of machine learning in the context of official statistics.‡

Keywords: Bias Correction · Misclassification Bias · Supervised Machine Learning · Classification · Official Statistics

†Corresponding authors: k.kloos@cbs.nl, q.a.meertens@uva.nl

‡The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors would like to thank Arnout van Delden and three anonymous referees for their useful comments on previous versions of this paper.

1 Introduction

Currently, many researchers in the field of official statistics are examining the potential of machine learning algorithms. A typical example is estimating the proportion of houses in the Netherlands having solar panels, by employing a machine learning algorithm trained to classify satellite images [3]. However, as long as the algorithm's predictions are not error-free, the estimate of the relative occurrence of a class, also known as the *base rate*, can be biased [17,18]. This fact is also intuitively clear: if the number of false positives does not equal the number of false negatives, then the estimate of the base rate is biased, even if the false positive rate and false negative rate are both small. The statistical bias that occurs when aggregating the predictions of a machine learning algorithm is referred to as *misclassification bias* [5].

Misclassification bias occurs in a broad range of applications, including official statistics [13], land cover mapping [12], political science [9,21], and epidemiology [8]. The objective in each of these applications is to minimize a loss function at the level of aggregated predictions, in contrast to minimizing a loss function at the level of individual predictions. Within the field of machine learning, learning with that objective is referred to as quantification learning; see [6] for a recent overview. In quantification learning, the idea is not to train a classifier at all, but to directly estimate the base rate from the feature distribution. A drawback of that approach is that relatively large training and test datasets are needed to optimize hyperparameters and to obtain accurate estimates of the accuracy of the prediction, respectively. In the applications referred to before, labelled data are often expensive to obtain and therefore scarce. Hence, in this paper, we focus on what is referred to as quantifiers based on corrected classifiers [6]. In short, it entails that we first aggregate predictions of classification algorithms and then correct the aggregates in order to reduce misclassification bias.

In the literature on measurement error, several methods have been proposed to reduce misclassification bias when aggregating categorical data that is prone to measurement error; see [11] for a technical discussion and [1] for a more recent overview. Based on that literature, we propose a total of five estimators for the base rate that can be derived from the confusion matrix of a classification algorithm. As reducing bias might increase variance, the estimators are evaluated by their mean squared error (MSE). To the best of our knowledge, for three of the five estimators, only asymptotic expressions for the MSE are ever presented in the literature. In this paper, we derive the expressions for the MSE for finite datasets. As a first step, we restrict ourselves to binary classification problems. Nonetheless, we believe that the same proof strategies may be used for multi-class classification problems. The expressions for the MSE enable a theoretical comparison of the five estimators for finite datasets. It allows us, for the first time, to make solid recommendations on how to employ classification algorithms in official statistics and other disciplines interested in aggregate statistics.

The remainder of the paper is organized as follows. First, in Section 2, the five estimators are formally introduced and the mathematical expressions for their MSEs are presented. The derivations are included in the appendix. Then, in

Section 3, the decision boundaries are numerically derived. We can indicate under which condition, like the sensitivity and specificity of the learning algorithm and the size of the test set, each of the estimators has the lowest MSE. Finally, in Section 4, we draw our main conclusion and discuss directions for future research.

2 Methods

Consider a *target population* of N objects and assume that the objects can be separated into two classes. One of the two classes is the *class of interest*. We refer to the relative occurrence of the class of interest in the target population as the *base rate* and we denote that parameter by α. In the example mentioned in Section 1, the objects are houses in the Netherlands and the two classes are whether or not the house has solar panels on the roof [3]. The class of interest is having solar panels and hence α indicates the relative frequency of houses in the country having solar panels.

We assume that the true classifications are only known for objects in a small simple random sample of the target population. In the applications that we consider, these classifications are obtained by manual inspection of the objects in that sample. Objects that belong to the class of interest receive class label 1, the other objects receive class label 0. Then, the sample is split randomly into a training set and a test set. As usual, the training set is used for model selection through cross-validation and is then used to train the selected model. We will consider the result of that part of the process as given. The test set is used to estimate the classification performance of the trained algorithm, which we will discuss in more detail below. Finally, the classification algorithm is applied on the entire target population (minus the small random sample, but we will neglect that small difference) resulting in a predicted label for each object.

As we will encounter in Subsection 2.2, simply computing the relative occurrence of objects predicted to belong to the class of interest will result in a biased estimate of α. That bias is referred to as *misclassification bias* [4]. In this section, five estimators for the base rate parameter α are formally introduced, many of which have been proposed decades ago; see [11] for an extensive discussion. We summarize the formulas for bias and variance that can be found in the literature and complement them with our own derivations.

In order to correct for misclassification bias, we need estimates of the algorithm's (mis)classification probabilities. Following [20], we assume that misclassifications are independent across objects and that the (mis)classification probabilities are the same for each object, conditional on their true class label. With this classification-error model in mind, we denote the probability that the algorithm predicts an object of class 0 correctly by p_{00} and we define p_{11} analogously. Observe that p_{11} and p_{00} correspond to the algorithm's sensitivity and specificity, respectively. The *confusion matrix* $\mathbf{P}$ is then defined as follows:

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

Table 1: Contingency tables for test set (left) and target population (right)

(a)

		Estimated class		
		0	1	Tot
True class	0	n_{00}	n_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+}
	Tot	n_{+0}	n_{+1}	n

(b)

		Estimated class		
		0	1	Tot
True class	0	N_{00}	N_{01}	N_{0+}
	1	N_{10}	N_{11}	N_{1+}
	Tot	N_{+0}	N_{+1}	N

The classification probabilities p_{00} and p_{11} are not known, but will be estimated using the test set. We write n for the size of the test set and introduce the notation n_{ij} and N_{ij} as depicted in Table 1. The classification probabilities are then estimated without bias by $\hat{p}_{00} = n_{00}/n_{0+}$ and $\hat{p}_{11} = n_{11}/n_{1+}$. (Here, the assumption is needed that the test set is a simple random sample from the target population.) Furthermore, the base rate α for the target population is defined formally as $\alpha = N_{1+}/N$.

Finally, we make the following technical assumptions. We assume that the algorithm is not perfect in predicting either of the classes, but that it is better than guessing for both of the classes, i.e., we assume that $0.5 < p_{ii} < 1$. Because the test set is a small (i.e., $n \ll N$) simple random sample from the population, n_{0+} may be assumed to follow a $Bin(n, \alpha)$-distribution, since α is considered fixed. Moreover, the classification-error model that we assume implies that the elements in the rows in Table 1, conditional on the corresponding row total, follow a binomial distribution as well, with the corresponding classification probability as success probability. For example, to name just two out of the eight entries, $n_{00} \mid n_{0+} \sim Bin(n_{0+}, p_{00})$ and $N_{10} \mid N_{1+} \sim Bin(N_{1+}, 1-p_{11})$. Last, the assumption $n \ll N$ justifies our ultimate technical assumption, which is that the estimators for the entries in $\mathbf{P}$ based on the test set on the one hand and estimators for α based only on the predicted class labels for the target population on the other hand, are independent random variables.

2.1 Baseline estimator - random sample

The baseline estimator for α is the proportion of data points in the test dataset for which the observed class label is equal to 1. The baseline estimator will be denoted by $\hat{\alpha}_a$. Under the assumptions discussed above, it is immediate that $\hat{\alpha}_a$ is an unbiased estimator for α, i.e.:

$$B\left[\hat{\alpha}_a\right] = 0. \tag{2}$$

Since we have assumed that the size n of the test dataset is much smaller than the size N of the population data, we may approximate the distribution of $n\hat{\alpha}_a$ by a binomial distribution with success probability α. The variance, and hence the MSE, of $\hat{\alpha}_a$ is then given by

$$MSE\left[\hat{\alpha}_a\right] = V\left[\hat{\alpha}_a\right] = \frac{\alpha(1-\alpha)}{n}. \tag{3}$$

This MSE will serve as the baseline value for the other estimators we discuss.

2.2 Classify and count

When applying a trained machine learning algorithm on new data, we may simply count the number of data points for which the predicted class equals 1. The resulting estimator for α, which we will denote by $\hat{\alpha}^*$, is referred to as the 'classify-and-count' estimator, see [6]. In general, the classify-and-count estimator is (strongly) biased, and has almost zero variance. More specifically,

$$E\left[\hat{\alpha}^*\right] = \alpha p_{11} + (1-\alpha)(1-p_{00}), \tag{4}$$

and hence

$$B\left[\hat{\alpha}^*\right] = \alpha(p_{11}-1) + (1-\alpha)(1-p_{00}), \tag{5}$$

which is zero only if the point (p_{00}, p_{11}) lies on the line through $(1-\alpha, \alpha)$ and $(1,1)$ in $\mathbb{R}^2$, as shown in [17]. The variance of the classify-and-count estimator is derived in [2] and equals

$$V\left[\hat{\alpha}^*\right] = \frac{\alpha p_{11}(1-p_{11}) + (1-\alpha)p_{00}(1-p_{00})}{N}. \tag{6}$$

If the population size N is large, the variance of $\hat{\alpha}^*$ is low. In some literature, this low variance is misinterpreted as high accuracy, by claiming intuitively that the large size of the dataset implies that the noise cancels out (cf. [16]). However, the nonzero bias is neglected in such arguments. Therefore, we are interested in the MSE because it considers both bias and variance. It equals

$$MSE\left[\hat{\alpha}^*\right] = \Big[\alpha(p_{11}-1) + (1-\alpha)(1-p_{00})\Big]^2 + O\left(\frac{1}{N}\right). \tag{7}$$

Here and below, the notation $O(1/x)$ indicates a remainder term that, for sufficiently large values of $x > 0$, is always contained inside an interval $(-C/x, C/x)$ for some constant $C > 0$; see, e.g., [19, p. 147]. Observe how, in general, the MSE does not converge to 0 as N tends to ∞.

2.3 Subtracting estimated bias

Knowing that the classify-and-count estimator $\hat{\alpha}^*$ is biased (see (5)), we may attempt to estimate that bias and subtract it from $\hat{\alpha}^*$. As briefly mentioned in [17], we may estimate that bias by the plug-in estimator, that is, we substitute the unknown quantities in Equation (5) by their estimates. More precisely, the bias is estimated as

$$\widehat{B}\left[\hat{\alpha}^*\right] = \hat{\alpha}^*(\hat{p}_{00} + \hat{p}_{11} - 2) + (1-\hat{p}_{00}), \tag{8}$$

in which the estimators $\hat{p}_{00}$ and $\hat{p}_{11}$ are based on the test dataset. The resulting estimator $\hat{\alpha}_b$ for α equals

$$\hat{\alpha}_b = \hat{\alpha}^* - \widehat{B}\left[\hat{\alpha}^*\right] = \hat{\alpha}^*(3 - \hat{p}_{00} - \hat{p}_{11}) - (1-\hat{p}_{00}). \tag{9}$$

To the best of our knowledge, the bias and variance of the estimator $\hat{\alpha}_b$ have not been published in the scientific literature. Therefore, we have derived both, up to terms of order $1/n^2$, yielding the following result.

Theorem 1. *The bias of $\hat{\alpha}_b$ as estimator for α is given by*

$$B\left[\hat{\alpha}_b\right] = (1-p_{00})(2-p_{00}-p_{11}) - \alpha(p_{00}+p_{11}-2)^2. \tag{10}$$

The variance of $\hat{\alpha}_b$ equals

$$\begin{aligned} V\left[\hat{\alpha}_b\right] = {} & \frac{\left[\alpha(p_{00}+p_{11}-1)-p_{00}\right]^2 p_{00}(1-p_{00})}{n(1-\alpha)}\left(1+\frac{\alpha}{n(1-\alpha)}\right) \\ & + \frac{\left[\alpha(p_{00}+p_{11}-1)+(1-p_{00})\right]^2 p_{11}(1-p_{11})}{n\alpha}\left(1+\frac{1-\alpha}{n\alpha}\right) \\ & + O\left(\max\left[\frac{1}{n^3},\frac{1}{N}\right]\right). \end{aligned} \tag{11}$$

Proof. See the Appendix.

In particular, Theorem 1 implies that $B\left[\hat{\alpha}_b\right] = (2-p_{00}-p_{11})B\left[\hat{\alpha}^*\right]$, compare Equations (10) and (5). Hence, $|B\left[\hat{\alpha}_b\right]| \leq |B\left[\hat{\alpha}^*\right]|$, because $1 < p_{00}+p_{11} < 2$.

2.4 Misclassification probabilities

Let $\mathbf{P}$ be the row-normalized confusion matrix of the machine learning algorithm that we have trained, as defined in (1). That is, entry p_{ij} is the probability that the algorithm predicts class j for a data point that belongs to class i. The probabilities p_{ij} are referred to as misclassification probabilities. In the binary setting, we write $\boldsymbol{\alpha}$ for the column vector $(1-\alpha,\alpha)^T$ (similarly for $\hat{\boldsymbol{\alpha}}^*$). Under the assumption that the probabilities p_{ij} are identical for each data point, we obtain the expression $E[\hat{\boldsymbol{\alpha}}^*] = \mathbf{P}^T\boldsymbol{\alpha}$. If the true values of all entries p_{ij} of $\mathbf{P}$ were known and if $p_{00}+p_{11} \neq 1$, then $\hat{\boldsymbol{\alpha}}_p = (\mathbf{P}^T)^{-1}\hat{\boldsymbol{\alpha}}^*$ would be an unbiased estimator for α. Using the plug-in estimator $\hat{\mathbf{P}}$ for $\mathbf{P}$, estimated on the test set, the following estimator for α is obtained:

$$\hat{\alpha}_p = \frac{\hat{\alpha}^* + \hat{p}_{00} - 1}{\hat{p}_{00}+\hat{p}_{11}-1}. \tag{12}$$

It is known that the estimator $\hat{\alpha}_p$ is consistent (asymptotically unbiased) for α, see [1]. In [7], the variance of this estimator is analysed for an arbitrary number of classes. For the binary case, a simple analytic expression for the bias and variance of $\hat{\alpha}_p$ for finite datasets has not been given, as far as we know. Therefore, we have derived the bias and variance for finite datasets, yielding the following result.

Theorem 2. *The bias of $\hat{\alpha}_p$ as estimator for α is given by*

$$B\left[\hat{\alpha}_p\right] = \frac{p_{00} - p_{11}}{n(p_{00} + p_{11} - 1)} + O\left(\frac{1}{n^2}\right). \tag{13}$$

The variance of $\hat{\alpha}_p$ is given by

$$V[\hat{\alpha}_p] = \frac{(1-\alpha)p_{00}(1-p_{00})\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + \alpha p_{11}(1-p_{11})\left[1 + \frac{1-\alpha}{n\alpha}\right]}{n(p_{00} + p_{11} - 1)^2} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right). \tag{14}$$

Proof. See the Appendix.

2.5 Calibration probabilities

Let $\mathbf{C}$ be the column-normalized confusion matrix of the machine learning algorithm that we have trained. That is, entry c_{ij} is the probability that the true class of a data point is j given that the algorithm has predicted class i. The probabilities c_{ij} are referred to as calibration probabilities [11]. The first element of the vector $\mathbf{C}\hat{\boldsymbol{\alpha}}^*$ is an unbiased estimator for $\boldsymbol{\alpha}$, if $\mathbf{C}$ is known.

Using the plug-in estimator $\hat{\mathbf{C}}$ for $\mathbf{C}$, which is estimated on the test dataset analogously to $\hat{\mathbf{P}}$, the following estimator $\hat{\alpha}_c$ for α is obtained:

$$\hat{\alpha}_c = \hat{\alpha}^* \frac{n_{11}}{n_{+1}} + (1 - \hat{\alpha}^*)\frac{n_{10}}{n_{+0}}, \tag{15}$$

in which each n_{ij} and n_{+j} should be considered as random variables. It has been shown that $\hat{\alpha}_c$ is a consistent estimator for α [1]. Under the assumptions we have made in this paper, it can be shown that $\hat{\alpha}_c$ is in fact an unbiased estimator for α. To the best of our knowledge, we are also the first to give an approximation (up to terms of order $1/n^2$) of the variance of $\hat{\alpha}_c$. Both results are summarized in the following theorem.

Theorem 3. *The calibration estimator $\hat{\alpha}_c$ is an unbiased estimator for α:*

$$B\left[\hat{\alpha}_c\right] = 0. \tag{16}$$

The variance of $\hat{\alpha}_c$ is equal to the following expression:

$$\begin{aligned} V(\hat{\alpha}_c) = &\left[\frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n} + \frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n^2}\right] \\ &\times \left[\frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\right)\right] \\ &+ \left[\frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n} + \frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n^2}\right] \\ &\times \left[\frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\right)\right] \\ &+ O\left(\max\left[\frac{1}{n^3}, \frac{1}{Nn}\right]\right). \end{aligned} \tag{17}$$

Proof. See the Appendix.

Hereby, the overview of the five estimators for α is complete. The expressions that we have derived for the bias and variance of these five estimators will now be used to compare the (root) mean squared error of the five estimators, both theoretically as well as by means of simulation studies.

3 Results

The aim of this section is to derive empirically which of the five estimators of α that we presented in Section 2 has the lowest MSE, and under which conditions. For a given population size N, the MSE of each estimator depends on four parameters (i.e, $\alpha, p_{00}, p_{11}, n$), so visualizations would have to be 5-dimensional. To reduce dimensions, we will first present a simulation study in which all four parameters are fixed. For the fixed parameter setting, the sampling distributions of the estimators are compared using boxplots. Second, we will fix several values of α and n and use plots to compare the MSE of the estimators for varying p_{00} and p_{11}. The latter analysis will already be sufficient in order to reach a final conclusion on which estimator has the lowest MSE.[‖]

3.1 Sampling distributions of the estimators

Here, we present two simple simulation studies to gain some intuition for the difference in the sampling distributions of the five estimators. In the first simulation study, we consider a class-balanced dataset, that is, $\alpha = 0.5$, with a small test dataset of size $n = 1000$, a large population dataset $N = 3 \times 10^5$ and a rather poor classifier having classification probabilities $p_{00} = 0.6$ and $p_{11} = 0.7$. We deliberately choose $p_{00} \neq p_{11}$, as otherwise the classify-and-count estimator $\hat{\alpha}^*$ would be unbiased: (p_{00}, p_{11}) would be on the line between $(1-\alpha, \alpha)$ and $(1, 1)$, see also Equation (5).

Table 2 summarizes the bias, variance and root mean squared error (RMSE), computed using the analytic approximations presented in Section 2. The classify-and-count estimator is highly biased and therefore it has a high RMSE, despite having the lowest variance of all estimators. The RMSE of the classify-and-count estimator can indeed be improved by subtracting an estimate of the bias ($\hat{\alpha}_b$). The subtraction reduces the absolute bias and only slightly increases the variance. A further bias reduction is obtained by the misclassification estimator $\hat{\alpha}_p$. However, inverting the row-normalized confusion matrix $\mathbf{P}$ (that is, the misclassification probabilities) for values of p_{00} and p_{11} close to $p_{00}+p_{11} = 1$ significantly

[‖]The results in this section have been obtained using the statistical software R. All visualizations have been implemented in a Shiny dashboard, which in addition includes interactive 3D-plots of the RMSE surface for each of the estimators. The code, together with the appendix, can be retrieved from `https://github.com/kevinkloos/Misclassification-Bias`.

increases the variance of the estimator, leading to the highest RMSE of all estimators considered. Finally, the calibration estimator $\hat{\alpha}_c$ is unbiased and has the lowest variance among the estimators that make use of the test dataset. In particular, note that the variance is also lower than that of the baseline estimator. In this example, the estimator based on the calibration probabilities has the lowest RMSE, and it is the only estimator with a lower RMSE than the baseline estimator $\hat{\alpha}_a$.

Table 2: A comparison of the bias, variance and RMSE of each of the five estimators for α, where $\alpha = 0.5$, $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$.

Estimator	*Symbol*	Bias $\times 10^{-2}$	Variance $\times 10^{-4}$	RMSE $\times 10^{-2}$
Baseline	$\hat{\alpha}_a$	0.000	2.500	1.581
Classify-and-count	$\hat{\alpha}^*$	5.000	0.000	5.000
Subtracted-bias	$\hat{\alpha}_b$	3.500	2.244	3.807
Misclassification	$\hat{\alpha}_p$	-0.033	25.025	5.003
Calibration	$\hat{\alpha}_c$	0.000	2.275	1.508

To gain insight in the sampling distribution of the estimators, in addition to the metrics presented in Table 2, we simulated a large number $R = 10000$ of confusion matrices for datasets of size $n = 1000$ and $N = 3 \times 10^5$. Each confusion matrix was created as follows. First, take a random draw from a $Bin(N, \alpha)$-distribution, resulting in a number N_{1+}. Then, take a random draw from a $Bin(N_{1+}, p_{11})$-distribution to obtain N_{11} and a random draw from a $Bin(N - N_{+1}, p_{00})$-distribution to obtain N_{00}. This computes the theoretical confusion matrix for the target population. Use this confusion matrix to draw a sample from a multivariate hypergeometric distribution, with its parameters from the drawn theoretical confusion matrix. These draws precisely give the number of true and false positives and negatives needed to fill a confusion matrix. Each confusion matrix can be used to compute the five estimators. Repeating this procedure $R = 10000$ times gave rise to the sampling distributions of the five estimators as presented in Figure 1. It nicely visualizes the bias and variance of the five estimators, supporting the results in Table 2. In addition, it shows that, due to the bias, the variances of the classify-and-count estimator $\hat{\alpha}^*$ and the subtracted-bias estimator $\hat{\alpha}_b$ cannot be used to obtain reliable confidence intervals for α.

In the second simulation study, we consider a highly imbalanced dataset, namely $\alpha = 0.98$. We again assume that the available test dataset has size $n = 1000$, but we assume a classifier having classification probabilities $p_{00} = 0.94$ and $p_{11} = 0.97$. Table 3 summarizes the bias, variance and RMSE of each of the estimators and Figure 2 shows the sampling distributions of each of the estima-

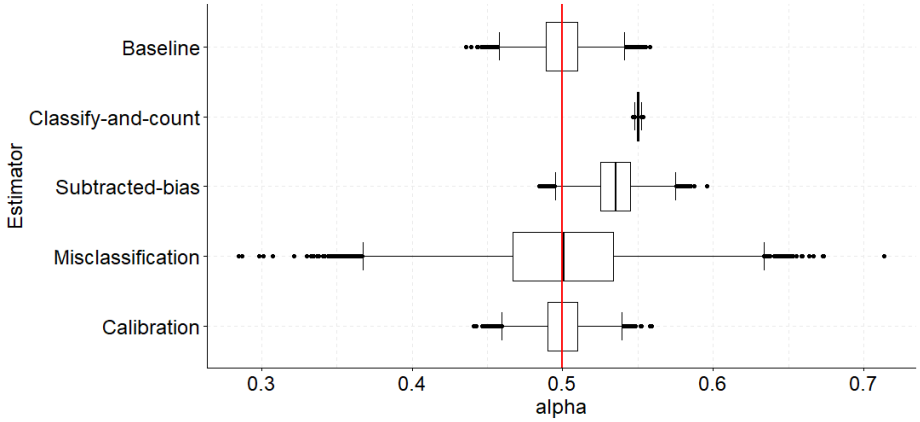

Fig. 1: The boxplots show the sampling distribution of the estimators for α, where $\alpha = 0.5$, $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$. The true value of α is highlighted by a vertical line.

tors. It can be noticed that subtracted-bias estimator and the misclassification estimator both have estimates of α that exceed 1. It is obvious that such values cannot occur in the population. For the method with the misclassification probabilities, this effect gets stronger when $p_{00} + p_{11}$ gets closer to 1. Furthermore, the baseline estimator performs well compared to the other estimators when the dataset is highly imbalanced: its RMSE is slightly higher than the RMSE of the method with calibration probabilities and much lower than the method with the misclassification probabilities. Finally, it is shown that the classify-and-count estimator is highly biased, even though p_{00} and p_{11} are both fairly close to 1.

Table 3: A comparison of the bias, variance and RMSE of each of the five estimators for α, where $\alpha = 0.98$, $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$.

Method	*Symbol*	Bias $\times 10^{-2}$	Variance $\times 10^{-5}$	RMSE $\times 10^{-3}$
Baseline	$\hat{\alpha}_a$	0.000	1.960	4.427
Classify-and-count	$\hat{\alpha}^*$	−2.820	0.000	28.200
Subtracted-bias	$\hat{\alpha}_b$	−0.254	3.377	6.342
Misclassification	$\hat{\alpha}_p$	−0.003	3.587	5.989
Calibration	$\hat{\alpha}_c$	0.000	1.289	3.591

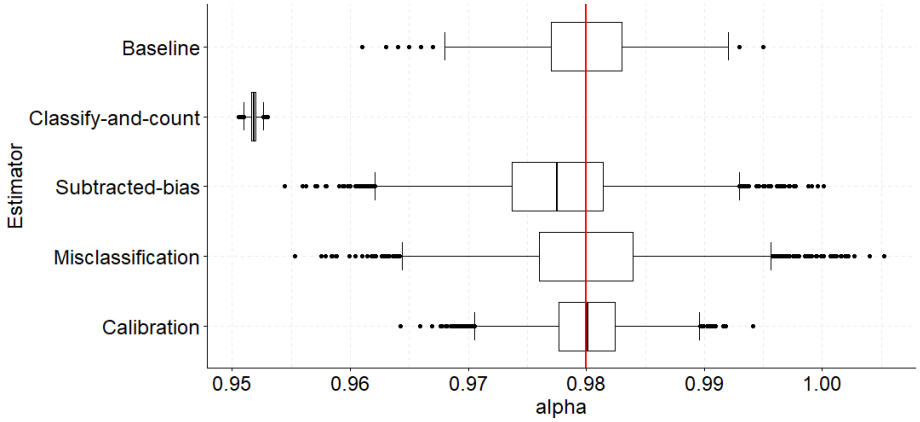

Fig. 2: The boxplots show the sampling distribution of the estimators for α, where $\alpha = 0.98$, $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$. The true value of α is highlighted by a vertical line.

3.2 Finding the optimal estimator

The aim of this subsection is to find the optimal estimator, i.e., the estimator with the lowest RMSE, for every combination of values of the parameters α, p_{00}, p_{11} and n. First, suppose that (p_{00}, p_{11}) is close to the line in the plane through the points $(1-\alpha, \alpha)$ and $(1, 1)$. As noted before, it implies that the classify-and-count estimator $\hat{\alpha}^*$ has low bias. Consequently, the subtracted-bias estimator $\hat{\alpha}_b$ has low bias as well. Thus, these two estimators will have the lowest RMSE in the described region, whose size decreases as n increases. Figure 3 visualizes the described region for $\alpha = 0.2$ and two different values of n. We remark that the biased estimators $\hat{\alpha}^*$ and $\hat{\alpha}_b$ perform worse (relative to the other estimators) when the sample size n of the test dataset increases. The biased methods, like Classify-and-count and Subtracted-bias, perform well when the classification probabilities are high for the largest group.

As we have seen in both Table 2 and Table 3, the calibration estimator $\hat{\alpha}_c$ competes with the baseline estimator in having the lowest RMSE. In general, the calibration estimator will have lower RMSE if the classification probabilities p_{00} and p_{11} are higher, while the baseline estimator does not depend on these classification probabilities. In a neighbourhood of $p_{00} = p_{11} = 0.5$, the baseline estimator will always have lower RMSE than the calibration estimator. However, for every α and n, there must exist a curve in the (p_{00}, p_{11})-plane beyond which the calibration estimator will have lower RMSE than the baseline estimator. The left-hand panels in Figure 4 show this curve for $\alpha = 0.2$ and two different values of n. For larger values of n, the curve where the calibration estimator performs better than the baseline estimator gets closer to $p_{00} = p_{11} = 0.5$ and therefore covers a larger area in the (p_{00}, p_{11})-plane.

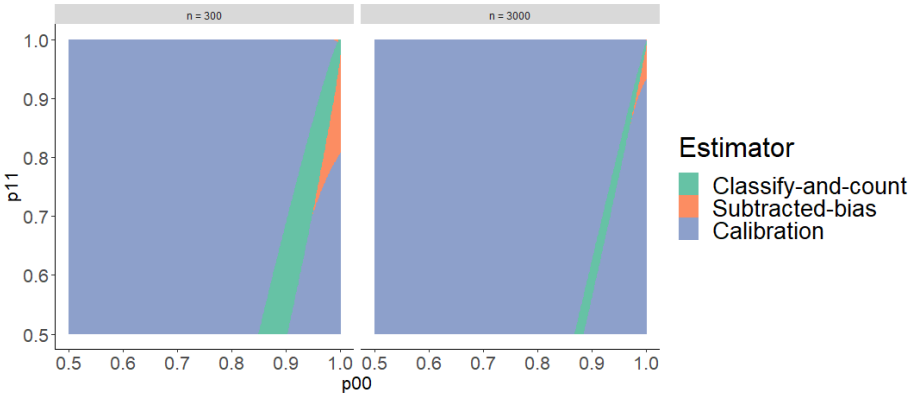

Fig. 3: For each coordinate (p_{00}, p_{11}), the depicted color indicates which estimator has the lowest RMSE, considering only the classify-and-count estimator (green), the subtracted-bias estimator (orange) and the calibration estimator (purple). In the left panel , we have set $\alpha = 0.2$ and $n = 300$, whereas $\alpha = 0.2$ and $n = 3000$ in the right panel. The red and green regions are smaller in the right panel, as the variance of the calibration estimator is decreasing in n, while the bias of the classify-and-count estimator and of the subtracted-bias estimator do not depend on n.

Table 2 and Table 3 have shown that the misclassification estimator only performs well if p_{00} and p_{11} are high, which is confirmed by the expression of the bias and variance: both have a singularity at $p_{00} + p_{11} = 1$, see Equations (13) and (14). The right-hand panels in Figure 4 show, for $\alpha = 0.2$ and two different values of n, the curve in the (p_{00}, p_{11})-plane beyond which the misclassification estimator has lower RMSE than the baseline estimator. Observe that an increase in the size n of the test dataset does not have much impact on the position of the curve. The reason is that the misclassification estimator has a singularity at $p_{00} = p_{11} = 0.5$. The shape of the curve also depends on the value of α. If $\alpha = 0.8$ instead of 0.2, the curves are line-symmetric in the line $p_{00} = p_{11}$. The curve is also line symmetric in $p_{00} = p_{11}$ for $\alpha = 0.5$. The area where the misclassification estimator performs better than the baseline estimator decreases when α gets closer towards 0 or 1. The main reason why this happens is that the variance of the baseline estimator decreases fast when α gets closer towards 0 or 1. Thus, the baseline estimator performs better than the misclassification estimator either if the classifier performs badly in general or performs badly in classifying the largest group.

The final analysis of this paper is to compare the calibration estimator and the misclassification estimator for high values of p_{00} and p_{11}. In Theorem 4 it is proven that, for all possible combinations of α and sufficiently large n, the MSE of the calibration estimator is consistently lower than that of the misclassification estimator.

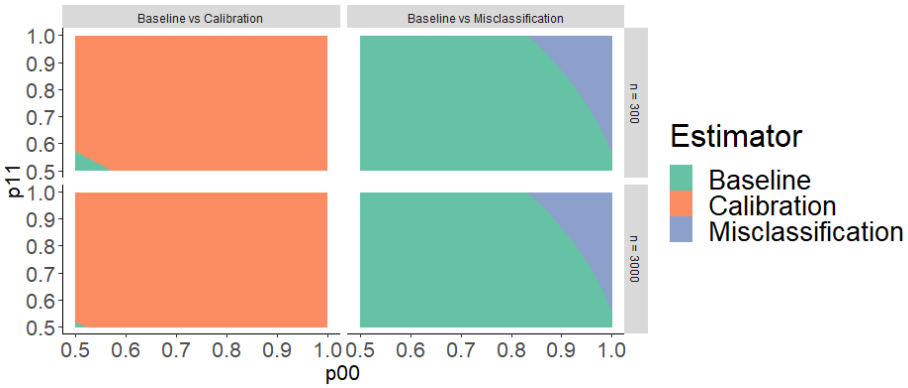

Fig. 4: For each coordinate (p_{00}, p_{11}), the depicted color indicates which estimate has the lowest RMSE, considering only the baseline estimator (green), the calibration estimator (orange) and the misclassification estimator (purple). The top-row panels consider $\alpha = 0.2$ and $n = 300$, while the bottom-row panels consider $\alpha = 0.2$ and $n = 3000$.

Theorem 4. *Let* $\widetilde{MSE}[\hat{\alpha}_p]$ *and* $\widetilde{MSE}[\hat{\alpha}_c]$ *denote the approximate mean squared errors, up to terms of order* $1/n$*, of the misclassification estimator and the calibration estimator, respectively. It holds that:*

$$\widetilde{MSE}[\hat{\alpha}_p] - \widetilde{MSE}[\hat{\alpha}_c] = \frac{\Big[(1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})\Big]^2}{(p_{00}+p_{11}-1)^2\beta(1-\beta)}, \tag{18}$$

in which $\beta := (1-\alpha)(1-p_{00}) + \alpha p_{11}$.

Proof. See the Appendix.

Thus, neglecting terms of order $1/n^2$ and higher, the result implies that the calibration estimator has a lower mean squared error than the misclassification estimator, except that both are equal if and only if $p_{00} = p_{11} = 1$. (Note that $0 < \beta < 1$.)

We do remark that the difference in MSE is large in particular for values of p_{00} and p_{11} close to $\frac{1}{2}$. More specifically, it diverges when $p_{00} + p_{11} \to 1$. It is the result of the misclassification estimator having a singularity at $p_{00} + p_{11} = 1$ (see Equation (14)), while the variance of the calibration estimator is bounded. An unpleasant consequence of the singularity at $p_{00} + p_{11} = 1$ is that, for fixed n and α, the probability that $\hat{\alpha}_p$ takes values outside the interval $[0, 1]$ increases as $p_{00} + p_{11} \to 1$; see [14] for a discussion and a possible solution.

4 Conclusion and Discussion

In this paper, we have studied the effect of classification errors on five estimators of the base rate parameter α that are obtained from machine learning algorithms.

In general, a straightforward classify-and-count estimator will lead to biased estimates and some form of bias correction should be considered. As reducing bias might increase variance, we evaluated the (root) mean squared error (MSE) of the five estimators, both theoretically as well as numerically.

From our results we may draw the following main (three-part) conclusion regarding which estimator for α has lowest mean squared error. First, when dealing with small test datasets and rather poor algorithms, that is p_{00} and p_{11} both close to 0.5, the baseline estimator $\hat{\alpha}_a$ has the lowest MSE. Second, when dealing with algorithms for which the classification probabilities p_{00} and p_{11} are in a small neighbourhood around the line $(p_{11} - 1)\alpha + (1 - p_{00})(1 - \alpha) = 0$ in the (p_{00}, p_{11})-plane, the classify-and-count estimator and the subtracted-bias estimator will have the lowest MSE. As the size of the test dataset increases, the size of that neighbourhood decreases. Third, in any other situation, the calibration estimator will have the lowest MSE. In practice, the test dataset will have to be used to determine which of the three scenarios applies to the data and the algorithm at hand. It is an additional estimation problem that we have not discussed in this paper.

We would like to close the paper by pointing out three interesting directions for future research. First, the results could be generalized to multi-class classification problems. The theoretical derivations of the bias and variance are more complicated and involve matrix-vector notation, but the proof strategy is similar. However, it is more challenging to compare the MSE of the five estimators visually in the multi-class case.

Second, the assumptions that we have made could be relaxed. In particular, a trained and implemented machine learning model is, in practice, often used over a longer period of time. A shift in the base rate parameter α, also known as prior probability shift [15], is then inevitable. Consequently, we may no longer assume that the conditional distribution of the class label given the features in the test dataset is similar to that in the population. It implies that the calibration estimator is no longer unbiased, which might have a significant effect on our main conclusion.

Third and finally, a combination of estimators might have a substantially lower MSE than that of the individual estimators separately. Therefore, it might be interesting to study different methods of model averaging applied to the problem of misclassification bias. It could be fruitful especially when the assumptions that we have made are relaxed.

References

1. Buonaccorsi, J.P.: Measurement Error: Models, Methods, and Applications. Chapman & Hall/CRC, Boca Raton, FL (2010)
2. Burger, J., Delden, A.v., Scholtus, S.: Sensitivity of Mixed-Source Statistics to Classification Errors. Journal of Official Statistics **31**(3), 489–506 (2015)
3. Curier, R., De Jong, T., Strauch, K., Cramer, K., Rosenski, N., Schartner, C., Debusschere, M., Ziemons, H., Iren, D., Bromuri, S.: Monitoring Spatial Sustainable

Development: Semi-Automated Analysis of Satellite and Aerial Images for Energy Transition and Sustainability Indicators. arXiv preprint arXiv:1810.04881 (2018)
4. Czaplewski, R.L.: Misclassification Bias in Areal Estimates. Photogrammetric Engineering and Remote Sensing. 58(2): 189-192 **58**(2), 189–192 (1992)
5. Czaplewski, R.L., Catts, G.P.: Calibration of Remotely Sensed Proportion or Area Estimates for Misclassification Error. Remote Sensing of Environment **39**(1), 29–43 (1992)
6. González, P., Castaño, A., Chawla, N.V., Coz, J.J.D.: A Review on Quantification Learning. ACM Computing Surveys **50**(5), 74:1–74:40 (2017)
7. Grassia, A., Sundberg, R.: Statistical Precision in the Calibration and Use of Sorting Machines and Other Classifiers. Technometrics **24**(2), 117–121 (1982)
8. Greenland, S.: Sensitivity Analysis and Bias Analysis. In: Ahrens, W., Pigeot, I. (eds.) Handbook of Epidemiology. Springer, New York, NY (2014)
9. Hopkins, D.J., King, G.: A Method of Automated Nonparametric Content Analysis for Social Science. American Journal of Political Science **54**(1), 229–247 (2010)
10. Knottnerus, P.: Sample Survey Theory: Some Pythagorean Perspectives. Springer Science & Business Media (2003)
11. Kuha, J., Skinner, C.J.: Categorical Data Analysis and Misclassification. In: Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., Trewin, D. (eds.) Survey Measurement and Process Quality, pp. 633–670. Wiley (Mar 1997)
12. Löw, F., Knöfel, P., Conrad, C.: Analysis of Uncertainty in Multi-Temporal Object-Based Classification. ISPRS Journal of Photogrammetry and Remote Sensing **105**, 91–106 (2015)
13. Meertens, Q.A., Diks, C.G.H., Herik, H.J.v.d., Takes, F.W.: A Data-Driven Supply-Side Approach for Estimating Cross-Border Internet Purchases Within the European Union. Journal of the Royal Statistical Society: Series A (Statistics in Society) **183**(1), 61–90 (2020)
14. Meertens, Q., Diks, C., van den Herik, H., Takes, F.: A Bayesian Approach for Accurate Classification-Based Aggregates. Society for Industrial and Applied Mathematics, Philadelphia, PA (2019)
15. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A Unifying View on Dataset Shift in Classification. Pattern Recognition **45**(1), 521–530 (2012)
16. O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, N.: From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media. Washington, DC (2010)
17. Scholtus, S., van Delden, A.: On the Accuracy of Estimators Based on a Binary Classifier (Feb 2020), Discussion Paper, Statistics Netherlands, The Hague
18. Schwartz, J.E.: The Neglected Problem of Measurement Error in Categorical Data. Sociological Methods & Research **13**(4), 435–466 (1985)
19. Strichartz, R.S.: The Way of Analysis. Jones & Bartlett Learning (2000)
20. Van Delden, A., Scholtus, S., Burger, J.: Accuracy of Mixed-Source Statistics as Affected by Classification Errors. Journal of Official Statistics **32**(3), 619–642 (2016)
21. Wiedemann, G.: Proportional Classification Revisited: Automatic Content Analysis of Political Manifestos Using Active Learning. Social Science Computer Review **37**(2), 135–159 (2019)