Statistical Journal of the IAOS 37 (2021) 1181–1196
DOI 10.3233/SJI-210885
IOS Press

A new generic method to improve machine learning applications in official statistics

Kevin Kloos
Intern Methodologist, Statistics Netherlands (CBS), https://www.cbs.nl/en-gb/, The Hague, The Netherlands
E-mail: k.kloos@fsw.leidenuniv.nl

Abstract. The use of machine learning algorithms at national statistical institutes has increased significantly over the past few years. Applications range from new imputation schemes to new statistical output based entirely on machine learning. The results are promising, but recent studies have shown that the use of machine learning in official statistics always introduces a bias, known as misclassification bias. Misclassification bias does not occur in traditional applications of machine learning and therefore it has received little attention in the academic literature. In earlier work, we have collected existing methods that are able to correct misclassification bias. We have compared their statistical properties, including bias, variance and mean squared error. In this paper, we present a new generic method to correct misclassification bias for time series and we derive its statistical properties. Moreover, we show numerically that it has a lower mean squared error than the existing alternatives in a wide variety of settings. We believe that our new method may improve machine learning applications in official statistics and we aspire that our work will stimulate further methodological research in this area.

Keywords: Misclassification bias, quantification, machine learning, official statistics, bias correction

1. Introduction

National statistical institutes (NSIs) currently apply many different types of machine learning algorithms. Classification algorithms are one of the most popular types of algorithms because publishing aggregate statistics of (sub)groups in a population is one of the main tasks of national statistical institutes. Classical examples of classification algorithms are logistic regression and linear discriminant analysis, but also new innovative algorithms have been introduced over the last decades, like additive models, decision trees and deep learning [1]. Classification algorithms are optimized to minimize the summed loss of individual units, such that each unit has a high probability to be classified correctly. However, classifying units individually can lead to biased results when generalizing these individual units to aggregate statistics, like a proportion of the population [2,3]. The cause of the biased results are imbalanced errors.

Before we show how generalizing units to aggregate statistics can lead to bias, we first emphasize the difference between a classifier and a quantifier. A classifier is a model that labels each unit to a class and a quantifier is a model that counts the number of units labelled to a class. Quantifiers can use classifiers in their model by counting the number of labels that the classifier has assigned to each class. Classifiers and quantifiers are imperfect because classification algorithms can mislabel some units. Each unit has a classification probability of being labelled correctly by the classification algorithm. A well-performing classifier has high classification probabilities for each labelled unit. A well-performing quantifier is not particularly defined by the number of mislabeled units, but by how the number of mislabeled units are distributed among the classes. In almost all cases, the number of mislabeled units among the classes don't cancel each other out and as a consequence, bias will occur. The bias that occurs from imbalanced classification errors is called *misclassification bias*.

Misclassification bias cannot simply be solved by improving the accuracy of the classification algorithm. Moreover, a more accurate classifier can increase mis-

classification bias. For example, classifier A with 10 false positives and 10 false negatives is a worse classifier than classifier B with 9 false positives and 5 false negatives. However, when aggregating the results of both classifiers to a quantification, classifier A turns out to have less misclassification bias than classifier B and is, therefore, a better quantifier. Classifier A has less misclassification bias than classifier B because the number of mislabeled units in classifier A are equally distributed among both classes, while the number of mislabeled units in classifier B are unequally distributed among the classes. Therefore, improving a classifier is not the solution to reduce misclassification bias [2,3].

We illustrate misclassification bias more extensively using an image-labelling example. The example shows us why using a standard approach for aggregating classifications from machine learning classifiers leads to problems. Suppose that a local government wants to estimate the number of houses in a certain area with solar panels on their rooftops. There is no register whether a house has a solar panel installation or not. It is an expensive and time-consuming task to manually label each rooftop, so the government decides to use satellite images combined with a classification algorithm to quickly label each house whether it has solar panels or not. Our target population consists of 10,000 houses, whereof 1,000 houses with solar panels and 9,000 without solar panels. Thus, the true proportion of houses with solar panels is 10%. The target variable is the proportion of houses with solar panels installation. Assume that the classifier can predict the rooftop images fairly accurate: 98% of the houses with solar panels are classified correctly (sensitivity) and 92% of the houses without solar panels are classified correctly (specificity). The machine learning algorithm classifies then 98% of the houses with solar panels and 8% of the houses without solar panels as houses with solar panels. This aggregates to $1{,}000 \times 0.98 + 9{,}000 \times 0.08 = 1{,}700$ houses classified as a house with solar panels installation by the machine learning classifier. Thus, we estimate the proportion of houses with solar panels as 17% instead of the true value of 10%. The difference between the true proportion and the estimated proportion of houses with solar panels is called misclassification bias, and as the example demonstrates it can occur even when the classifier can predict every individual label with high accuracy.

In the literature, several corrections methods exist to reduce misclassification bias of the proportion of units labelled to the class of interest, i.e. the *base rate*. We compared statistical properties of the five most-used correction methods in a previous paper [4]. The correction methods contain information from the target population and a test set, see Table 1. The target population consists of N units that are labelled by a classification algorithm where we want to estimate the base rate from. The true labels are unknown in the target population, only the estimated labels are available. Therefore, the confusion matrix of the target population cannot be constructed in practice; only the column totals (white cells) in Table 1a are known. We construct a test set to get more information on the accuracy of the classifier. The test set consists of $n \ll N$ randomly sampled units from the target population that are both labelled by a classification algorithm and a human classifier. Therefore, we can construct a confusion matrix from the test set, see Table 1b, which contains information about the classification probabilities and the true base rate. In contrast to the target population, all the cells in the test set are known. The correction methods used in [4] exclusively contain information from the test set and the target population whereof closed-form equations of the mean square error (MSE) for each correction method could be computed. [4] concluded that the so-called calibration estimator works the best in general (more information in the next section).

Table 1
Confusion matrices of the target population and test set. Grey values are unknown in practice

(a) Target population

	Estimate		
True	Class 0	Class 1	Total
Class 0	N_{00}	N_{01}	N_{0+}
Class 1	N_{10}	N_{11}	N_{1+}
Total	N_{+0}	N_{+1}	N

(b) Test set

	Estimate		
True	Class 0	Class 1	Total
Class 0	n_{00}	n_{01}	n_{0+}
Class 1	n_{10}	n_{11}	n_{1+}
Total	n_{+0}	n_{+1}	n

However, the result from that paper does not generalize to time series. In other words, the results could not be applied for populations where the base rate changes over time. The target populations that are interesting for national statistical institutes, where we produce statistics on a monthly, quarterly or annual basis, change from period to period. The solar panel case is a good quantification example for time series: households can place solar panels on their roofs or displace them during a certain period. Moreover, the proportion of houses with solar panels is an interesting statistic concerning

the government's aims of renewable energy. The drift that occurs when the target population changes over time, is called concept drift [5]. In this paper, we assume a special case of concept drift called prior probability shift. Prior probability shift assumes that the base rate of a target population changes over time, but that the classification probabilities of units conditioned on their true label remain constant over time [6].

The most effective, but most costly and time-consuming solution to deal with prior probability shift, is to construct a new test set for each period. A more cost-efficient solution is to construct a test set and use the same test from period to period. As a consequence, we then cannot assume that a test set is a simple random sample of the target population when the base rate changes over time. Therefore, new expressions for bias and variance are needed to evaluate the MSE of the five correction methods. These expressions were previously computed by [7]. They concluded that none of those estimators performs consistently well under prior probability shift.

The main contribution of this paper is a new generic method to correct for misclassification bias when dealing with prior probability shift. We will refer to the resulting estimator as the *mixed estimator* because it combines the strengths of two existing estimators. We will derive (approximate) closed-form expressions for the bias and variance of the mixed estimator. Moreover, we will numerically compare the mixed estimator's MSE with the classical methods.

The remainder of the paper is organized as follows: in Section 2, we introduce the problem and assumptions and we recap the properties of the original correction methods. Section 3 introduces the mixed estimator. Moreover, we will compare the mixed estimator with the original correction methods. Section 4 contains a discussion and conclusion of this paper.

2. Model under prior probability shift

In this section, we introduce the quantifier under prior probability shift. We use the same mathematical approach as in [4] and therefore use the same parameters and assumptions. Before we dive into the mathematical expressions, we briefly discuss the terminology used in the later sections. The target population has N units which belong to one of two classes, either class 0 or class 1. Our parameter of interest is the proportion of units that belong to class 1 in the target population, denoted as α. Similarly to [8], we assume that the underlying classifier has a probability of p_{00} to correctly classify an object of class 0 and a probability of p_{11} to correctly classify an object of class 1. These classification probabilities are unknown in practice, so we randomly sample a test set of size n from the target population. In the test set, both the true labels and the estimated labels are known. Then, we can make an estimate for p_{00} ($\hat{p}_{00} = \frac{n_{00}}{n_{0+}}$) and for p_{11} ($\hat{p}_{11} = \frac{n_{11}}{n_{1+}}$) with the test set. We assume that a binary classification algorithm has been trained that correctly classifies a data point that belongs to class $i \in \{0, 1\}$ with probability $p_{ii} > 0.5$, independently across all data points. In addition, we assume that a test set of size $n \ll N$ is available and that it can be considered a simple random sample from the population. Finally, we assume that the classify-and-count estimator $\hat{\alpha}^*$ is distributed independently of $\hat{p}_{00}$ and $\hat{p}_{11}$, which is reasonable (at least as an approximation) when $n \ll N$.

In this paper, we allow that the base rate can change over time. In other words, we allow for a nonzero prior probability shift. Therefore, we introduce the following notation. First, we need to distinguish a target population U at time 0 from a target population U' at time t. We can then define α' as the base rate of the target population U'. Moreover, the classification probabilities of the target population are equal for U and U', so the new base rate α' is the only new parameter in this paper. Therefore, α denotes the base rate of target population U, whereof the test set is constructed.

Before we describe the differences between [4] and [7], we briefly introduce the correction methods. First, the baseline estimator ($\hat{\alpha}_a$) computes the proportion of units in the test set that belong to class 1. Second, the classify-and-count estimator ($\hat{\alpha}^\star$) computes the proportion of units that are classified by the machine learning algorithm to class 1 in the target population. This is the naive estimator where we simply count the number of units that belong to class 1 according to the algorithm. Third, the subtracted-bias ($\hat{\alpha}_b$) estimator first estimates the bias of the classify-and-count estimator by estimating classification probabilities in the test set. Then, we compute the subtracted-bias estimator by subtracting this estimated bias from the classify-and-count estimator. Fourth, the misclassification estimator ($\hat{\alpha}_p$) multiplies the inverted row-normalised test set by the classify-and-count estimator. Last, the calibration estimator ($\hat{\alpha}_c$) multiplies the column-normalised test set by the classify-and-count estimator. An overview of the equations can be found in Table 2, as well as how these estimators perform in terms of bias and variance. The baseline, misclassification and calibration estimator

are all (asymptotically) unbiased where the calibration estimator performs the best in general [7]. computed closed-form expressions of the bias and variance under prior probability shift. First, prior probability shift does not affect the mean square error of the classify-and-count estimator and subtracted-bias estimator majorly. The classify-and-count estimator does not use any information from the test set, so a different base rate does not affect this estimator. The subtracted-bias estimator only uses the test set to estimate the classification probabilities and therefore the estimator is not affected largely by the shift. Obviously, the baseline estimator cannot be used when the target population follows a different distribution than the test set. The misclassification estimator remains asymptotically unbiased, but the calibration estimator is unfortunately biased under prior probability shift. This is in contrast to the situation under a fixed base rate, where the calibration estimator is unbiased. Therefore, the misclassification estimator remains the only estimator that is (asymptotically) unbiased, see Table 3. According to [4] and [7], the misclassification estimator has a high variance when the classification probabilities are low. On the other hand, the variance of the misclassification estimator only changes slightly under prior probability shift. All in all, none of these correction methods have a consistently low MSE. In the next section, we will introduce a new estimator that performs better than the five original correction methods.

3. Mixed estimator

In this section, we introduce a new estimator: the mixed estimator. The mixed estimator is a combination between the misclassification estimator [9] and the calibration estimator [10]. In [4,7], we found that the calibration estimator is unbiased under a fixed base rate, but becomes biased under prior probability shift. The misclassification estimator has a higher variance, but the MSE remains fairly stable under prior probability shift. These two properties can be combined: as an initial starting point, we take the calibration estimator $\hat{\alpha}_c$ at time 0, but we add the difference between the misclassification estimator at time t ($\hat{\alpha}'_p$) and time 0 ($\hat{\alpha}_p$). Therefore, the expression for the mixed estimator is:

$$\begin{aligned}\hat{\alpha}'_m &= \hat{\alpha}_c + [\hat{\alpha}'_p - \hat{\alpha}_p] \\ &= \frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^*) + \frac{n_{11}}{n_{+1}}\hat{\alpha}^* + \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}. \quad (1)\end{aligned}$$

To the best of our knowledge, this is the first paper where the mixed estimator is introduced. Therefore, the closed-form expressions for bias and variance that we have derived are new as well.

Lemma 1. The variance of the estimator $\hat{p}_{11}$ for p_{11} estimated on the test set is given by

$$V(\hat{p}_{11}) = \frac{p_{11}(1-p_{11})}{n\alpha}\left[1+\frac{1-\alpha}{n\alpha}\right] + O\left(\frac{1}{n^3}\right). \quad (2)$$

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

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

Moreover, $\hat{p}_{11}$ and $\hat{p}_{00}$ are uncorrelated: $C(\hat{p}_{11}, \hat{p}_{00}) = 0$.

Theorem 1. The mixed estimator $\hat{\alpha}'_m$ is a biased, but consistent, estimator for $\alpha' \neq \alpha$:

$$B[\hat{\alpha}'_m] = \frac{(\alpha'-\alpha)(V(\hat{p}_{00})+V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} + O\left(\frac{1}{n^2}\right). \quad (4)$$

The variance of $\hat{\alpha}'_m$ is equal to:

$$\begin{aligned}&V(\hat{\alpha}'_m) \\ &= \frac{\alpha p_{11}}{n} \times \left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\right) \\ &+ \frac{(1-\alpha)p_{00}}{n} \times \left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\right) \\ &+ (\alpha'-\alpha)^2 \times \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^2} + (\alpha'-\alpha) \\ &\times \left[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)}\right. \\ &\left. - \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)}\right] \\ &+ O\left(\frac{1}{n^2}\right). \quad (5)\end{aligned}$$

Proof: See Appendix.

From Theorem 1, we see that the mixed estimator has a bias of $O\left(\frac{1}{n}\right)$. Therefore, the mixed estimator is slightly biased but consistent. The variance function is complex, but we can see that the variance will be larger when the difference between α' and α increases. To obtain a better overview of this mixed estimator, we will perform three simulation studies. Each simulation study

Table 2
Overview of the estimators without prior probability shift from [4]

Estimator	Equation	Bias	Variance
Baseline	$\hat{\alpha}_a = \frac{n_{1+}}{n}$	No	Large
Classify-and-count	$\hat{\alpha}^\star = \frac{N_{+1}}{N}$	Large	Very low
Subtracted-bias	$\hat{\alpha}_b = \hat{p}_{11}\hat{\alpha}^\star + (1-\hat{p}_{00})(1-\hat{\alpha}^\star)$	Medium	Low
Misclassification	$\hat{\alpha}_p = \frac{\hat{\alpha}^\star + \hat{p}_{00} - 1}{\hat{p}_{00} + \hat{p}_{11} - 1}$	Very low	Large
Calibration	$\hat{\alpha}_c = \frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^\star) + \frac{n_{11}}{n_{+1}}\hat{\alpha}^\star$	No	Medium

Table 3
Overview of the estimators under prior probability shift from [7]

Estimator	Equation	Bias	Variance
Baseline	$\hat{\alpha}'_a = \frac{n_{1+}}{n}$	Large	Large
Classify-and-count	$(\hat{\alpha}^\star)' = \frac{N'_{+1}}{N}$	Large	Very low
Subtracted-bias	$\hat{\alpha}'_b = \hat{p}_{11}(\hat{\alpha}^\star)' + (1-\hat{p}_{00})(1-(\hat{\alpha}^\star)')$	Medium	Low
Misclassification	$\hat{\alpha}'_p = \frac{(\hat{\alpha}^\star)' + \hat{p}_{00} - 1}{\hat{p}_{00} + \hat{p}_{11} - 1}$	Very low	Large
Calibration	$\hat{\alpha}'_c = \frac{n_{10}}{n_{+0}}(1-(\hat{\alpha}^\star)') + \frac{n_{11}}{n_{+1}}(\hat{\alpha}^\star)'$	Medium	Medium

consists of $B = 10{,}000$ estimates for each correction method. First, we create fixed target populations given $\alpha, \alpha', p_{00}, p_{11}$ and N: a population U at time 0 and a population U' at time t. Then, we sample $B = 10{,}000$ test sets of size n from population U and apply the estimators and equations from Table 3 on each test set.

In the first simulation study, we consider a class-balanced dataset ($\alpha = 0.5$), with a small test set of size $n = 1{,}000$, a large population dataset of $N = 3 \times 10^5$ and a rather poor classifier having classification probabilities $p_{00} = 0.6$ and $p_{11} = 0.7$. From Fig. 1, we can see that the mixed estimator is in general a stable estimator with a low amount of bias and much less variance than the misclassification estimator. However, the variance of the mixed estimator tends to increase when the difference between α' and α gets larger, which is in line with the observations in the previous paragraph. The mixed estimator performs much better than the misclassification estimator and the calibration estimator: it has almost no bias and has much less variance than the misclassification estimator.

A situation where the mixed estimator does not work as well as expected, can be found in Fig. 2. We specify the following parameters: $p_{00} = 0.94$, $p_{11} = 0.97$, $\alpha = 0.98$, $n = 1{,}000$, $N = 3 \times 10^5$ and $B = 10{,}000$. The misclassification estimator tends to have more extreme outliers when the difference between α' and α increases. This affects the mixed estimator in terms of variance. Furthermore, the mixed estimator can predict values outside the $[0, 1]$-interval. We cannot encounter these values in practice and it is, therefore, a problem that we obtain these estimates. It seems that this problem occurs less often for the mixed estimator than for the misclassification estimator so the mixed estimator performs still better than the misclassification estimator and the calibration estimator individually. Finally, we can observe that the variance of the mixed estimator is always lower than the variance of the misclassification estimator. Despite the outliers, it is still the best estimator out of the three.

In the first two simulation studies, the misclassification estimator did not work properly, and we showed values of α' that are close to α. It is also interesting to see what happens when the misclassification estimator has a low MSE for α and what happens when α' differs substantially from α. We perform a simulation study with $\alpha = 0.75$, $p_{00} = 0.85$, $p_{11} = 0.90$, $n = 1{,}000$, $N = 3 \times 10^5$ and $B = 10{,}000$, shown in Fig. 3. We observe that the distribution of the mixed estimator is similar to the distribution of the misclassification estimator. The reason behind this is that the misclassification estimator performs similarly to the calibration estimator at time 0. However, the figures and the numbers show that the mixed estimator still performs consistently better than the misclassification estimator.

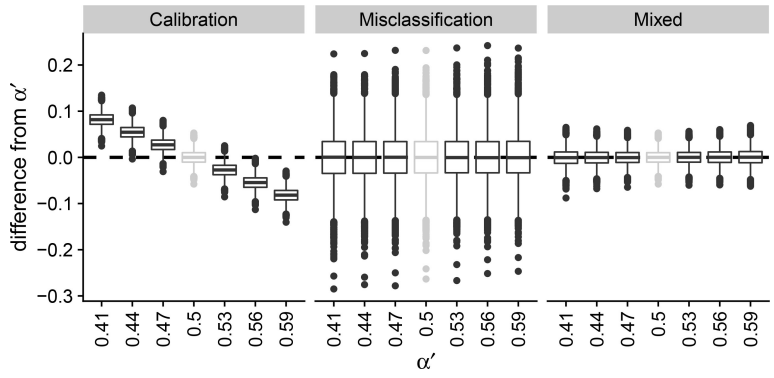

Fig. 1. Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given an initial base rate $\alpha = 0.5$ (grey) and different values of α' (black). The test set is sampled from the target population with the initial base rate. The x-axis shows the different base rates and the y-axis shows the distribution of the difference from the new base rate α'. All the parameters: $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1{,}000$ and $N = 3 \times 10^5$, $B = 10{,}000$.

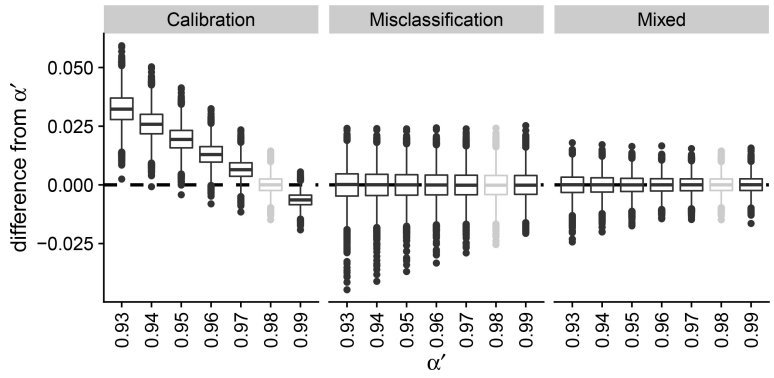

Fig. 2. Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given an initial base rate $\alpha = 0.98$ (grey) and different values of α' (black). The test set is sampled from the target population with the initial base rate. The x-axis shows the different base rates and the y-axis shows the distribution of the difference from the new base rate α'. All the parameters: $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1{,}000$ $N = 3 \times 10^5$ and $B = 10{,}000$.

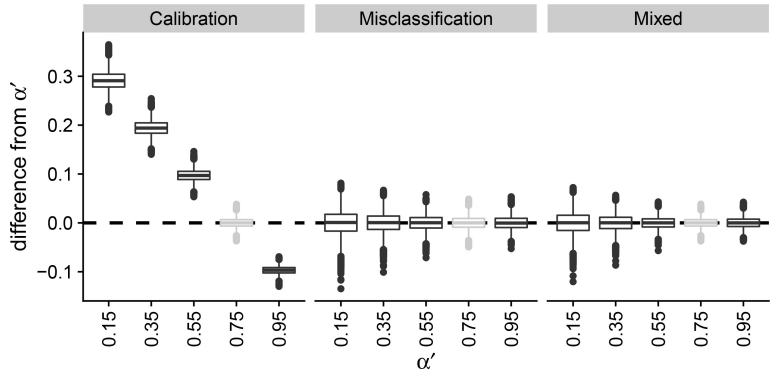

Fig. 3. Simulation study to observe the change in prediction error under concept drift using boxplots. The calibration, misclassification and mixed estimator are compared given an initial base rate $\alpha = 0.75$ (grey) and different values of α' (black). The test set is sampled from the target population with the initial base rate. The x-axis shows the different base rates and the y-axis shows the distribution of the difference from the new base rate α'. All the parameters: $p_{00} = 0.85$, $p_{11} = 0.90$, $n = 1{,}000$, $N = 3 \times 10^5$ and $B = 10{,}000$.

4. Conclusion and discussion

We conclude that our mixed estimator outperforms the estimators currently available in the academic literature. The mixed estimator has less bias than the calibration estimator and less variance than the calibration estimator. The mixed estimator performs much better than the calibration estimator and the misclassification estimator when the variance of the misclassification estimator is large but consistent over time. Our results show that the mixed estimator outperforms both the calibration estimator and the misclassification estimator in any dataset and for any classification algorithm used.

Even though that the new mixed estimator performs better than the original correction methods, we still believe that the correction methods might be improved further. We could construct a new estimator by combining biased, but invariant correction methods. New research directions lay in combining the correction methods in such a way that both bias and variance of the new estimator will be consistently low.

The estimator could be extended for correction methods that can predict more than two classes. The downside is that the number of parameters increases quadratically and the quality measure should be adapted for multiple classes. A possible solution is to further elaborate the simulation studies, instead of computing closed-form mathematical expressions. A final extension that we recommend is allowing the classification probabilities to differ between the units within a group, see [8].

With this paper, we hope that we raised awareness that aggregating outcomes of machine learning algorithms can be very inaccurate, even if the algorithms have a high prediction accuracy. Furthermore, this paper is an addition to the scientific literature on the theory of misclassification bias. Finally, we proposed a new generic method that can be used by NSIs to improve machine learning applications within official statistics.

References

[1] Friedman JH, Hastie T, Tibshirani R, et al. The elements of statistical learning. vol. 1. Springer, New York; 2001.

[2] Schwarz JE. The neglected problem of measurement error in categorical data. Sociological Methods & Research. 1985.

[3] Scholtus S, van Delden A. On the accuracy of estimators based on a binary classifier. 2020; 202006. Discussion Paper, Statistics Netherlands, The Hague.

[4] Kloos K, Meertens QA, Scholtus S, Karch JD. Comparing correction methods to reduce misclassification bias. in: Artificial Intelligence and Machine Learning. Cham: Springer International Publishing; Baratchi M, Cao L, Kosters WA, Lijffijt J, van Rijn JN, Takes FW, eds, 2021; pp. 64-90.

[5] Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F. Characterizing concept drift. Data Mining and Knowledge Discovery. 2016; 30(4): 964-994.

[6] Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern recognition. 2012; 45(1): 521-530.

[7] Meertens QA, Diks CGH, Van Den Herik HJ, Takes FW. Understanding the output quality of official statistics that are based on machine learning algorithms; 2021.

[8] van Delden A, Scholtus S, Burger J. Accuracy of mixed-source statistics as affected by classification errors. Journal of Official Statistics. 2016; 32(3): 619-642.

[9] Buonaccorsi JP. Measurement error: Models, methods, and applications. Boca Raton, FL: Chapman & Hall/CRC; 2010.

[10] Kuha J, Skinner CJ. Categorical data analysis and misclassification. in: Survey Measurement and Process Quality. Wiley; Lyberg LE, Biemer PP, Collins M, de Leeuw ED, Dippo C, Schwarz N, et al., eds, 1997; pp. 633-670.

[11] Knottnerus P. Sample survey theory: Some pythagorean perspectives. Springer Science & Business Media; 2003.

Appendix

This appendix contains the proofs of the theorems presented in the paper entitled: *A new generic method to improve machine learning applications in official statistics*. Recall that we have assumed a population of size N in which a fraction $\alpha := N_{1+}/N$ belongs to the class of interest, referred to as the class labelled as 1. We assume that a binary classification algorithm has been trained that correctly classifies a data point that belongs to class $i \in \{0, 1\}$ with probability $p_{ii} > 0.5$, independently across all data points. In addition, we assume that a test set of size $n \ll N$ is available and that it can be considered a simple random sample from the population. The classification probabilities p_{00} and p_{11} are estimated on that test set by row-normalizing the confusion matrix of the test set. Finally, we assume that the classify-and-count estimator $\hat{\alpha}^*$ is distributed independently of $\hat{p}_{00}$ and $\hat{p}_{11}$, which is reasonable (at least as an approximation) when $n \ll N$.

It may be noted that the estimated probabilities $\hat{p}_{11}$ and $\hat{p}_{00}$ cannot be computed if $n_{1+} = 0$ or $n_{0+} = 0$. Similarly, the calibration probabilities c_{11} and c_{00} cannot be estimated if $n_{+1} = 0$ or $n_{+0} = 0$. We assume here that these events occur with negligible probability. This will be true when n is sufficiently large so that $n\alpha \gg 1$ and $n(1-\alpha) \gg 1$.

Preliminaries

Many of the proofs presented in this appendix rely on the following two mathematical results. First, we will use univariate and bivariate Taylor series to approximate the expectation of non-linear functions of random variables. That is, to estimate $E[f(X)]$ and $E[g(X, Y)]$ for sufficiently differentiable functions f and g, we will insert the Taylor series for f and g at $x_0 = E[X]$ and $y_0 = E[Y]$ up to terms of order 2 and utilize the linearity of the expectation. Second, we will use the following conditional variance decomposition for the variance of a random variable X:

$$V(X) = E[V(X|Y)] + V(E[X|Y]). \tag{6}$$

The conditional variance decomposition follows from the tower property of conditional expectations [11]. Before we prove the theorems presented in the paper, we begin by proving Lemma 1.

Proof of Lemma 1 We approximate the variance of $\hat{p}_{00}$ using the conditional variance decomposition and a second-order Taylor series, as follows:

$$\begin{aligned}
V(\hat{p}_{00}) &= V\left(\frac{n_{00}}{n_{0+}}\right) \\
&= E_{n_{0+}}\left[V\left(\frac{n_{00}}{n_{0+}}\middle| n_{0+}\right)\right] + V_{n_{0+}}\left[E\left(\frac{n_{00}}{n_{0+}}\middle| n_{0+}\right)\right] \\
&= E_{n_{0+}}\left[\frac{1}{n_{0+}^2}V(n_{00}|n_{0+})\right] + V_{n_{0+}}\left[\frac{1}{n_{0+}}E(n_{00}|n_{0+})\right] \\
&= E_{n_{0+}}\left[\frac{n_{0+}p_{00}(1-p_{00})}{n_{0+}^2}\right] + V_{n_{0+}}\left[\frac{n_{0+}p_{00}}{n_{0+}}\right] \\
&= E_{n_{0+}}\left[\frac{1}{n_{0+}}\right] p_{00}(1-p_{00}) \\
&= \left[\frac{1}{E[n_{0+}]} + \frac{1}{2}\frac{2}{E[n_{0+}]^3} \times V[n_{0+}]\right] p_{00}(1-p_{00}) + O\left(\frac{1}{n^3}\right) \\
&= \frac{p_{00}(1-p_{00})}{E[n_{0+}]}\left[1 + \frac{V[n_{0+}]}{E[n_{0+}]^2}\right] + O\left(\frac{1}{n^3}\right) \\
&= \frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + O\left(\frac{1}{n^3}\right).
\end{aligned}$$

The variance of $\hat{p}_{11}$ is approximated in the exact same way.

Finally, to evaluate $C(\hat{p}_{11}, \hat{p}_{00})$ we use the analogue of Eq. (6) for covariances:

$$
\begin{aligned}
C(\hat{p}_{11}, \hat{p}_{00}) &= C\left(\frac{n_{11}}{n_{1+}}, \frac{n_{00}}{n_{0+}}\right) \\
&= E_{n_{1+},n_{0+}}\left[C\left(\frac{n_{11}}{n_{1+}}, \frac{n_{00}}{n_{0+}}\middle| n_{1+}, n_{0+}\right)\right] \\
&\quad + C_{n_{1+},n_{0+}}\left[E\left(\frac{n_{11}}{n_{1+}}\middle| n_{1+}, n_{0+}\right), E\left(\frac{n_{00}}{n_{0+}}\middle| n_{1+}, n_{0+}\right)\right] \\
&= E_{n_{1+},n_{0+}}\left[\frac{1}{n_{1+}n_{0+}} C(n_{11}, n_{00}|n_{1+}, n_{0+})\right] \\
&\quad + C_{n_{1+},n_{0+}}\left[\frac{1}{n_{1+}} E(n_{11}|n_{1+}), \frac{1}{n_{0+}} E(n_{00}|n_{0+})\right].
\end{aligned}
$$

The second term is zero as before. The first term also vanishes because, conditional on the row totals n_{1+} and n_{0+}, the counts n_{11} and n_{00} follow independent binomial distributions, so $C(n_{11}, n_{00}|n_{1+}, n_{0+}) = 0$.

Note: in the remainder of this appendix, we will not add explicit subscripts to expectations and variances when their meaning is unambiguous.

Mixed estimator

In this section, we will prove the bias and the variance of the mixed estimator under concept drift. The mixed estimator is dependent on the calibration estimator at time 0, the misclassification estimator on time 0 and the misclassification estimator on time t.

Proof of Theorem 1 First, we will make a proof for the bias of the Mixed Estimator. The expression for the Mixed Estimator is:

$$
\hat{\alpha}'_m = \hat{\alpha}_c + (\hat{\alpha}'_p - \hat{\alpha}_p) = \hat{\alpha}_c + [(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}. \tag{7}
$$

The bias is defined as the difference between the expected value of the estimator minus the true value of the target variable:

$$
B[\hat{\alpha}'_m] = E[\hat{\alpha}'_m] - \alpha' \tag{8}
$$

Using Eq. (7), we can write out the expected value of the mixed estimator.

$$
\begin{aligned}
E[\hat{\alpha}'_m] &= E\left[\hat{\alpha}_c + [(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] \\
&= E[\hat{\alpha}_c] + E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right]
\end{aligned} \tag{9}
$$

From [4], we already know that:

$$
E[\hat{\alpha}_c] = E[\hat{\alpha}_c|\hat{\alpha}^*] = \alpha. \tag{10}
$$

$E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]$ can be computed by conditioning on the Classify-and-count estimators $(\hat{\alpha}')^*$ and $\hat{\alpha}^*$.

$$
\begin{aligned}
E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right] &= E\left[E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\middle| (\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] \\
&= E\left[((\hat{\alpha}')^* - \hat{\alpha}^*) \times E\left[\frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\middle| (\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] \\
&= E\left[((\hat{\alpha}')^* - \hat{\alpha}^*) \times E\left[\frac{1}{\hat{p}_{00} + \hat{p}_{11} - 1}\right]\right]
\end{aligned} \tag{11}
$$

From [4], we used Taylor Series to approximate the expected value of $\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}$.

$$E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = \frac{1}{p_{00}+p_{11}-1} + \frac{V(\hat{p}_{00})+V\hat{p}_{11})}{(p_{00}+p_{11}-1)^3} + O(n^{-2}) \tag{12}$$

Now it only remains to calculate the expected values of the classify-and-count estimators.

$$E[(\hat{\alpha}')^* - \hat{\alpha}^*] = E[(\hat{\alpha}')^*] - E[\hat{\alpha}^*] \tag{13}$$

$$E[(\hat{\alpha}')^*] = \alpha' p_{11} + (1-\alpha')(1-p_{00}) = \alpha'(p_{00}+p_{11}-1) + (1-p_{00}) \tag{14}$$

$$E[\hat{\alpha}^*] = \alpha p_{11} + (1-\alpha)(1-p_{00}) = \alpha(p_{00}+p_{11}-1) + (1-p_{00}) \tag{15}$$

Combining these expressions, $E[(\hat{\alpha}')^* - \hat{\alpha}^*]$ can be simplified towards the following expression.

$$E[(\hat{\alpha}')^* - \hat{\alpha}^*] = (\alpha'-\alpha)(p_{00}+p_{11}-1) \tag{16}$$

Combining Eqs (12) and (16) gives the expression that should be in the big expectation of Eq. (11).

$$\begin{aligned}
E\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= E\left[((\hat{\alpha}')^* - \hat{\alpha}^*) \times E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]\right] \\
&= E[(\hat{\alpha}')^* - \hat{\alpha}^*] \times E\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\
&= (\alpha'-\alpha)(p_{00}+p_{11}-1) \times \left[\frac{1}{p_{00}+p_{11}-1} + \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^3}\right] + O(n^{-2}) \\
&= \alpha' - \alpha + \frac{(\alpha'-\alpha)(V(\hat{p}_{00})+V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} + O(n^{-2})
\end{aligned} \tag{17}$$

Finalizing the proof given Eqs (8), (10) and (17).

$$\begin{aligned}
B[\hat{\alpha}'_m] &= E[\hat{\alpha}'_m] - \alpha' \\
&= \alpha + \alpha' - \alpha + \frac{(\alpha'-\alpha)(V(\hat{p}_{00})+V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} - \alpha' + O(n^{-2}) \\
&= \frac{(\alpha'-\alpha)(V(\hat{p}_{00})+V(\hat{p}_{11}))}{(p_{00}+p_{11}-1)^2} + O(n^{-2})
\end{aligned} \tag{18}$$

Now it only remains to prove the variance of the mixed estimator. Recall that the mixed estimator can be written as

$$\hat{\alpha}'_m = \hat{\alpha}_c + [(\hat{\alpha}')^* - \hat{\alpha}^*] \times \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}. \tag{19}$$

It clearly follows from Eq. (19) that the variance of this mixed estimator can be written as

$$V[\alpha'_m] = V[\hat{\alpha}_c] + V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] + 2C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right]. \tag{20}$$

From [4], we already know that the variance of the calibration estimator is equal to

$$\begin{aligned}
V(\hat{\alpha}_c) = &\left[\frac{(1-\alpha)(1-p_{00})+\alpha p_{11}}{n} + \frac{(1-\alpha)p_{00}+\alpha(1-p_{11})}{n^2}\right] \\
&\times \left[\frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\right)\right] \\
&+ \left[\frac{(1-\alpha)p_{00}+\alpha(1-p_{11})}{n} + \frac{(1-\alpha)(1-p_{00})+\alpha p_{11}}{n^2}\right] \\
&\times \left[\frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\right)\right] \\
&+ O\left(\max\left[\frac{1}{n^3}, \frac{1}{Nn}\right]\right).
\end{aligned} \tag{21}$$

The second term in Eq. (20) makes use of previous assumptions in this paper. We can say that $\hat{p}_{00}$ and $\hat{p}_{11}$ are independent of our Classify-and-count estimators $\hat{\alpha}^*$ and $(\hat{\alpha}')^*$. Furthermore, a well-known result on variances states that for two independent random variables A and B, it holds that $V(AB) = E[A]^2V(B)+E[B]^2V(A)+V(A)V(B)$. Combining these statements gives

$$V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = [E((\hat{\alpha}')^* - \hat{\alpha}^*)]^2 V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] + \left[E\left(\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right)\right]^2 V[(\hat{\alpha}')^* - \hat{\alpha}^*] + V[(\hat{\alpha}')^* - \hat{\alpha}^*]V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]. \tag{22}$$

Assuming that $N \gg n$, we can make the statement that $V[(\hat{\alpha}')^* - \hat{\alpha}^*]$ is of $O\left(\frac{1}{N}\right)$.

$$V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = [E((\hat{\alpha}')^* - \hat{\alpha}^*)]^2 V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] + O\left(\frac{1}{N}\right) \tag{23}$$

The expected value of the differences between the classify-and-count estimators is already computed in Eq. (16) and the variance term in Eq. (23) is already proven in [4]. This eases the derivation of the second term in Eq. (20).

$$V\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = (\alpha'-\alpha)^2 \times \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^2} + O\left(\max\left[\frac{1}{N}, \frac{1}{n^2}\right]\right) \tag{24}$$

Thus it remains to evaluate the covariance term in Eq. (20). By conditioning on the classify-and-count estimators $\hat{\alpha}^*$ and $(\hat{\alpha}')^*$, we obtain:

$$C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = E\left[C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\middle|(\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] + C\left[E[\hat{\alpha}_c|(\hat{\alpha}')^*, \hat{\alpha}^*], E\left[\frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\middle|(\hat{\alpha}')^*, \hat{\alpha}^*\right]\right] \tag{25}$$

It can be proven that the second term of Eq. (25) is equal to zero. In Eq/ 10, we see that the expectation of the calibration estimator, given classify-and-count estimators, is equal to α. This is a constant and the covariance with a constant is equal to zero. Therefore, the covariance term can also be written as:

$$C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}^\star)' - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = E\left[C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}^\star)' - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\middle|(\hat{\alpha}^\star)', \hat{\alpha}^*\right]\right]. \tag{26}$$

We can derive an expression for the inner covariance, which is written as

$$C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}^\star)' - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\middle|(\hat{\alpha}^\star)', \hat{\alpha}^*\right] = [(\hat{\alpha}^\star)' - \hat{\alpha}^*]C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\middle|\hat{\alpha}^*\right]. \tag{27}$$

The terms in Eq. (27) can be written in terms of the test set $(n_{00}, n_{01}, n_{10}, n_{11})$. This eases the computation further on. Note that the elements of this test set do not depend on the classify-and-count estimator $\hat{\alpha}^*$.

$$\begin{aligned} C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\middle|\hat{\alpha}^*\right] &= C\left[\frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^*) + \frac{n_{11}}{n_{+1}}\hat{\alpha}^*, \frac{1}{\frac{n_{00}}{n_{0+}}+\frac{n_{11}}{n_{1+}}-1}\middle|\hat{\alpha}^*\right] \\ &= C\left[\frac{n_{10}}{n_{+0}}(1-\hat{\alpha}^*) + \frac{n_{11}}{n_{+1}}\hat{\alpha}^*, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle|\hat{\alpha}^*\right] \\ &= (1-\hat{\alpha}^*)C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] + \hat{\alpha}^* C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] \end{aligned} \tag{28}$$

We are able to evaluate both covariance terms with the same methods. We can condition on one of the row totals. Note that the other row total is also fixed, because we work with binary classifiers ($n_{1+} = n - n_{0+}$). Furthermore, we are able to write as many variables as possible in terms of n_{0+} and n_{1+}. This helps with the Taylor Series that we apply to approximate the covariances.

$$C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right]$$

$$
\begin{aligned}
&= E\left[C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right]\right] + C\left[E\left[\frac{n_{10}}{n_{+0}}\middle| n_{1+}\right], E\left[\frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right]\right] \\
&= E\left[C\left[\frac{n_{1+}-n_{11}}{n_{1+}+n_{00}-n_{11}}, \frac{n_{0+}n_{1+}}{n_{0+}n_{11}+n_{1+}n_{00}-n_{0+}n_{1+}}\middle| n_{1+}\right]\right] \\
&\quad + C\left[E\left[\frac{n_{1+}-n_{11}}{n_{1+}+n_{00}-n_{11}}\middle| n_{1+}\right], E\left[\frac{n_{0+}n_{1+}}{n_{0+}n_{11}+n_{1+}n_{00}-n_{0+}n_{1+}}\middle| n_{1+}\right]\right] \qquad (29)
\end{aligned}
$$

While we condition on the row totals, the other variables in the covariance functions are n_{00} and n_{11}. Say

$$\frac{n_{10}}{n_{+0}} = f(n_{00}, n_{11}) \text{ and } \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}} = g(n_{00}, n_{11}),$$

with

$$f(x,y) = \frac{n_{1+}-y}{n_{1+}+x-y} \qquad (30)$$

$$g(x,y) = \frac{n_{0+}n_{1+}}{n_{1+}x+n_{0+}y-n_{0+}n_{1+}} \qquad (31)$$

we are able to compute first-order Taylor series approximations for these terms to obtain an approximation for $C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right]$.

$$\frac{\partial f}{\partial x} = \frac{(n_{1+}+x-y)\cdot 0 - (n_{1+}-y)\cdot 1}{(n_{1+}+x-y)^2} = \frac{y-n_{1+}}{(n_{1+}+x-y)^2} \qquad (32)$$

$$\frac{\partial f}{\partial y} = \frac{(n_{1+}+x-y)\cdot -1 - (n_{1+}-y)\cdot -1}{(n_{1+}+x-y)^2} = \frac{-x}{(n_{1+}+x-y)^2} \qquad (33)$$

$$\frac{\partial g}{\partial x} = \frac{-(n_{0+}n_{1+})n_{1+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} = \frac{-n_{1+}^2 n_{0+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} \qquad (34)$$

$$\frac{\partial g}{\partial y} = \frac{-(n_{0+}n_{1+})n_{0+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} = \frac{-n_{0+}^2 n_{1+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^2} \qquad (35)$$

The approximation can be made with substituting $x = E[n_{00}|n_{1+}]$ and $y = E[n_{11}|n_{1+}]$ and applying the approximation rules for covariance. Given that n_{00} and n_{11} are independent from each other given the row totals, we can cross out $C(n_{00}, n_{11})$.

$$
\begin{aligned}
C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right] \approx{}& \frac{E[n_{11}|n_{1+}]-n_{1+}}{(n_{1+}+E[n_{00}|n_{1+}]-E[n_{11}|n_{1+}])^2} \\
&\times \frac{-n_{1+}^2 n_{0+}}{(n_{0+}E[n_{11}|n_{1+}]+n_{1+}E[n_{00}|n_{1+}]-n_{0+}n_{1+})^2} V(n_{00}|n_{1+}) \\
&+ \frac{-E[n_{00}|n_{1+}]}{(n_{1+}+E[n_{00}|n_{1+}]-E[n_{11}|n_{1+}])^2} \\
&\times \frac{-n_{0+}^2 n_{1+}}{(n_{0+}E[n_{11}|n_{1+}]+n_{1+}E[n_{00}|n_{1+}]-n_{0+}n_{1+})^2} V(n_{11}|n_{1+}) \qquad (36)
\end{aligned}
$$

In order to use this approximation, we can use the following properties:

$$E(n_{00}|n_{1+}) = n_{0+}p_{00}$$

$$V(n_{00}|n_{1+}) = n_{0+}p_{00}(1-p_{00})$$

$$E(n_{11}|n_{1+}) = n_{1+}p_{11}$$

$$V(n_{11}|n_{1+}) = n_{1+}p_{11}(1-p_{11})$$

Substituting these elements gives

$$C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right] \approx \frac{(n_{1+}p_{11})-n_{1+}}{(n_{1+}+n_{0+}p_{00}-n_{1+}p_{11})^2} \times \frac{-n_{1+}^2 n_{0+}}{(n_{0+}(n_{1+}p_{11})+n_{1+}(n_{0+}p_{00})-n_{0+}n_{1+})^2} n_{0+}p_{00}(1-p_{00}) + \frac{-n_{0+}p_{00}}{(n_{1+}+n_{0+}p_{00}-n_{1+}p_{11})^2} \times \frac{-n_{0+}^2 n_{1+}}{(n_{0+}(n_{1+}p_{11})+n_{1+}(n_{0+}p_{00})-n_{0+}n_{1+})^2} n_{1+}p_{11}(1-p_{11}). \quad (37)$$

This expression simplifies to

$$C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right] \approx \frac{n_{1+}p_{00}(1-p_{00})(1-p_{11})+n_{0+}p_{00}p_{11}(1-p_{11})}{(n_{1+}+n_{0+}p_{00}-n_{1+}p_{11})^2(p_{00}+p_{11}-1)^2} \quad (38)$$

Now that the inner covariance of Eq. (29) is computed, we can move on and calculate the inner expectations of Eq. (29). This can be done with a second-order Taylor series approximation.

$$\frac{\partial^2 f}{\partial x^2} = 2 \times \frac{n_{1+}-y}{(n_{1+}+x-y)^3} \quad (39)$$

$$\frac{\partial^2 f}{\partial y^2} = 2 \times \frac{-x}{(n_{1+}+x-y)^3} \quad (40)$$

$$\frac{\partial^2 g}{\partial x^2} = 2 \times \frac{n_{1+}^3 n_{0+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^3} \quad (41)$$

$$\frac{\partial^2 g}{\partial y^2} = 2 \times \frac{n_{0+}^3 n_{1+}}{(n_{0+}y+n_{1+}x-n_{0+}n_{1+})^3} \quad (42)$$

Applying the Taylor rules for approximating an expected value and substituting $x = E[n_{00}|n_{1+}]$ and $y = E[n_{11}|n_{1+}]$ into the equations gives:

$$E\left[\frac{n_{10}}{n_{+0}}\middle| n_{1+}\right] \approx \frac{n_{1+}-E[n_{11}|n_{1+}]}{n_{1+}+E[n_{00}|n_{1+}]-E[n_{11}|n_{1+}]} + \frac{n_{1+}-E[n_{11}|n_{1+}]}{(n_{1+}+E[n_{00}|n_{1+}]-E[n_{11}|n_{1+}])^3} V[n_{00}|n_{1+}] - \frac{E[n_{00}|n_{1+}]}{(n_{1+}+E[n_{00}|n_{1+}]-E[n_{11}|n_{1+}])^3} V[n_{11}|n_{1+}] \quad (43)$$

$$= \frac{n_{1+}-n_{1+}p_{11}}{n_{1+}+n_{0+}p_{00}-n_{1+}p_{11}} + \frac{n_{1+}-n_{1+}p_{11}}{(n_{1+}+n_{0+}p_{00}-n_{1+}p_{11})^3} n_{0+}p_{00}(1-p_{00}) - \frac{n_{0+}p_{00}}{(n_{1+}+n_{0+}p_{00}-n_{1+}p_{11})^3} n_{1+}p_{11}(1-p_{11}) \quad (44)$$

$$= \frac{n_{1+}(1-p_{11})}{n_{1+}+n_{0+}p_{00}-n_{1+}p_{11}} + \frac{n_{0+}n_{1+}p_{00}(p_{11}-1)(p_{00}+p_{11}-1)}{(n_{1+}+n_{0+}p_{00}-n_{1+}p_{11})^3} \quad (45)$$

$$E\left[\frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right] \approx \frac{n_{0+}n_{1+}}{n_{0+}E[n_{11}|n_{1+}]+n_{1+}E[n_{00}|n_{1+}]-n_{0+}n_{1+}} + \frac{n_{1+}^3 n_{0+}}{(n_{0+}E[n_{11}|n_{1+}]+n_{1+}E[n_{00}|n_{1+}]-n_{0+}n_{1+})^3} V[n_{00}|n_{1+}] + \frac{n_{0+}^3 n_{1+}}{(n_{0+}E[n_{11}|n_{1+}]+n_{1+}E[n_{00}|n_{1+}]-n_{0+}n_{1+})^3} V[n_{11}|n_{1+}] \quad (46)$$

$$
\begin{aligned}
&= \frac{n_{0+}n_{1+}}{n_{0+}n_{1+}p_{11} + n_{1+}n_{0+}p_{00} - n_{0+}n_{1+}} \\
&\quad + \frac{n_{1+}^3 n_{0+}}{(n_{0+}n_{1+}p_{11} + n_{1+}n_{0+}p_{00} - n_{0+}n_{1+})^3} n_{0+}p_{00}(1-p_{00}) \\
&\quad + \frac{n_{0+}^3 n_{1+}}{(n_{0+}n_{1+}p_{11} + n_{1+}n_{0+}p_{00} - n_{0+}n_{1+})^3} n_{1+}p_{11}(1-p_{11}) \qquad (47) \\
&= \frac{1}{p_{00}+p_{11}-1} + \frac{n_{1+}p_{00}(1-p_{00}) + n_{0+}p_{11}(1-p_{11})}{(n_{0+}n_{1+})(p_{00}+p_{11}-1)^3} \qquad (48)
\end{aligned}
$$

The next step is computing the outer expectation and the outer covariance of Eq. (29). The outer expectation can be approximated with a zero-order Taylor series.

$$
\begin{aligned}
E\left[C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right]\right] &\approx \frac{n\alpha p_{00}(1-p_{00})(1-p_{11}) + n(1-\alpha)p_{00}p_{11}(1-p_{11})}{(n\alpha + n(1-\alpha)p_{00} - n\alpha p_{11})^2(p_{00}+p_{11}-1)^2} \\
&= \frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00} - \alpha(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)^2} \qquad (49)
\end{aligned}
$$

Furthermore, it can be proven that the outer covariance of the two expectations is of $O\left(\frac{1}{n^2}\right)$ and can therefore be neglected in Eq. (29). In general, we can say that

$$
C[f(X), g(X)] \approx f'(E[X]) \times g'(E[X]) \times V(X) \qquad (50)
$$

Let $f(x)$ and $g(x)$ be the expectations of Eqs (45) and (48), with $x = n_{1+}$. Taking the derivative with respect to x gives:

$$
\begin{aligned}
f(x) &= \frac{x(1-p_{11})}{x + (n-x)p_{00} - xp_{11}} + \frac{(n-x)xp_{00}(p_{11}-1)(p_{00}+p_{11}-1)}{(x+(n-x)p_{00}-xp_{11})^3} \\
f'(x) &= \frac{np_{00}(p_{11}-1)}{(np_{00} - x(p_{00}+p_{11}-1))^2} \\
&\quad + \frac{[p_{00}(1-p_{11})(p_{00}+p_{11}-1)][(2x-n) + 3(x^2-nx)(np_{00} - x(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)}{(np_{00} - x(p_{00}+p_{11}-1)^6}
\end{aligned} \qquad (51)
$$

$$
\begin{aligned}
g(x) &= \frac{1}{p_{00}+p_{11}-1} + \frac{xp_{00}(1-p_{00}) + (n-x)p_{11}(1-p_{11})}{((n-x)x)(p_{00}+p_{11}-1)^3} \\
g'(x) &= \frac{(nx-x^2)(p_{00}(1-p_{00}) - p_{11}(1-p_{11})) + (2x-n)[xp_{00}(1-p_{00}) + (n-x)p_{11}(1-p_{11})])}{(nx-x^2)^2(p_{00}+p_{11}-1)^3} \qquad (52)
\end{aligned}
$$

If we substitute $x = E[n_{1+}] = n\alpha$ in the derivatives, we obtain the following expressions:

$$
\begin{aligned}
f'(E[n_{1+}]) &= \frac{p_{00}(p_{11}-1)}{n((1-\alpha)p_{00} + \alpha(1-p_{11}))^2} + p_{00}(1-p_{11})(p_{00}+p_{11}-1) \\
&\quad \times \frac{n(2\alpha-1) + 3n^4(1-\alpha)((1-\alpha)p_{00} + \alpha(1-p_{11}))^2(p_{00}+p_{11}-1)}{n^6((1-\alpha)p_{00} + \alpha(1-p_{11}))^6} \qquad (53) \\
g'(E[n_{1+}]) &= \frac{(\alpha-\alpha^2)(p_{00}(1-p_{00}) - p_{11}(1-p_{11}) + (2\alpha-1)(\alpha p_{00}(1-p_{00}) + (1-\alpha)p_{11}(1-p_{11}))}{n^2(\alpha-\alpha^2)(p_{00}+p_{11}-1)^3} \qquad (54)
\end{aligned}
$$

It can be clearly seen that $f'(E[n_{1+}]) = O\left(\frac{1}{n}\right)$, $g'(E[n_{1+}]) = O\left(\frac{1}{n^2}\right)$ and that $V(n_{1+}) = O(n)$. Therefore, the whole covariance term is small enough to be negligible ($O\left(\frac{1}{n}\right) \cdot O\left(\frac{1}{n^2}\right) \cdot O(n) = O\left(\frac{1}{n^2}\right)$, see Eq. (50)) and that the covariance term can be written as:

$$
C\left[\frac{n_{10}}{n_{+0}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] \approx \frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00} - \alpha(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)^2}. \qquad (55)
$$

Similarly, $C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right]$ can be computed. First, $C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{+1}\right]$ can be computed with a first-order Taylor series approximation. Because we condition on the row-totals, we rewrite $\frac{n_{11}}{n_{+1}}$ as

$$\frac{n_{11}}{n_{+1}} = \frac{n_{11}}{n - n_{00} - n_{10}} = \frac{n_{11}}{n - n_{00} - (n_{1+} - n_{11})} = \frac{n_{11}}{n_{0+} - n_{00} + n_{11}}$$

and make a function dependent on $x = n_{00}$ and $y = n_{11}$, which we can derive.

$$h(x,y) = \frac{y}{n_{0+} - x + y}$$

$$\frac{\partial h}{\partial x} = \frac{y}{(n_{0+} - x + y)^2} \tag{56}$$

$$\frac{\partial h}{\partial y} = \frac{n_{0+} - x}{(n_{0+} - x + y)^2} \tag{57}$$

Accordingly, we can borrow the expectations from the previous covariance term. Therefore we end up with the following term:

$$\begin{aligned} C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right] \approx & \frac{n_{1+}p_{11}}{(n_{0+}(1-p_{00}) + n_{1+}p_{11})^2} \\ & \times \frac{-n_{1+}^2 n_{0+}}{(n_{0+}(n_{1+}p_{11}) + n_{1+}(n_{0+}p_{00}) - n_{0+}n_{1+})^2} n_{0+}p_{00}(1-p_{00}) \\ & + \frac{n_{0+}(1-p_{00})}{(n_{0+}(1-p_{00}) + n_{1+}p_{11})^2} \\ & \times \frac{-n_{0+}^2 n_{1+}}{(n_{0+}(n_{1+}p_{11}) + n_{1+}(n_{0+}p_{00}) - n_{0+}n_{1+})^2} n_{1+}p_{11}(1-p_{11}). \end{aligned} \tag{58}$$

This simplifies to:

$$C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right] \approx -\frac{n_{1+}p_{00}(1-p_{00})p_{11} + n_{0+}(1-p_{00})p_{11}(1-p_{11})}{(n_{0+}(1-p_{00}) + n_{1+}p_{11})^2(p_{00}+p_{11}-1)^2} \tag{59}$$

The next step is computing the expected value of this expression.

$$\begin{aligned} E\left[C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\middle| n_{1+}\right]\right] &\approx -\frac{n\alpha p_{00}(1-p_{00})p_{11} + n(1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{(n(1-\alpha)(1-p_{00}) + n\alpha p_{11})^2(p_{00}+p_{11}-1)^2} \\ &= -\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00}) + \alpha p_{11})^2(p_{00}+p_{11}-1)^2} \end{aligned} \tag{60}$$

The covariance between the expectations is again of a negligible low order, so the covariance term can be written as:

$$C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{0+}n_{1+}}{n_{00}n_{11}-n_{01}n_{10}}\right] \approx -\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00}) + \alpha p_{11})^2(p_{00}+p_{11}-1)^2} \tag{61}$$

Now that we have obtained the two conditional covariance in Eqs (55) and (61), we can substitute these terms in Eq. (28).

$$\begin{aligned} C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\middle| (\hat{\alpha}')^*, \hat{\alpha}^*\right] \approx & (1-\hat{\alpha}^*) \times \frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n((1-\alpha)p_{00} + \alpha(1-p_{11}))^2(p_{00}+p_{11}-1)^2} \\ & -\hat{\alpha}^* \times \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00}) + \alpha p_{11})^2(p_{00}+p_{11}-1)^2} \end{aligned} \tag{62}$$

Combining Eqs (26), (27) and (62), we can compute $C[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}]$ by taking the expected value of the difference between the Classify-and-count estimators multiplied by the expected value of Eq. (62). Note that the first part of both denominators are equal to respectively the expected value of $(1-\hat{\alpha}^*)$ and $\hat{\alpha}^*$ squared.

$$
\begin{aligned}
C\left[\hat{\alpha}_c, \frac{(\hat{\alpha}')^* - \hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\right] &= E\left[[(\hat{\alpha}')^* - \hat{\alpha}^*] C\left[\hat{\alpha}_c, \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\middle|\hat{\alpha}^*\right]\right] \\
&\approx E\Bigg[[(\hat{\alpha}')^* - \hat{\alpha}^*]\Bigg[(1-\hat{\alpha}^*)\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))^2(p_{00}+p_{11}-1)^2} \\
&\quad -\hat{\alpha}^* \times \frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})^2(p_{00}+p_{11}-1)^2}\Bigg]\Bigg] \\
&\approx E\Bigg[[(\hat{\alpha}')^* - \hat{\alpha}^*]\Bigg[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)^2} \\
&\quad -\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)^2}\Bigg]\Bigg] \\
&= [(\alpha')-\alpha](p_{00}+p_{11}-1)\Bigg[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)^2} \\
&\quad -\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)^2}\Bigg] \\
&= [(\alpha')-\alpha]\Bigg[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)} \\
&\quad -\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)}\Bigg]
\end{aligned} \tag{63}
$$

Combining all elements gives the total variance of the mixed estimator.

$$
\begin{aligned}
V(\hat{\alpha}'_m) = &\frac{\alpha p_{11}}{n} \times \left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}}\right) \\
&+\frac{(1-\alpha)p_{00}}{n} \times \left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})}\right) + (\alpha'-\alpha)^2 \times \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^2} \\
&+(\alpha'-\alpha) \times \Bigg[\frac{\alpha p_{00}(1-p_{00})(1-p_{11}) + (1-\alpha)p_{00}p_{11}(1-p_{11})}{n(p_{00}-\alpha(p_{00}+p_{11}-1))(p_{00}+p_{11}-1)} \\
&-\frac{\alpha p_{00}(1-p_{00})p_{11} + (1-\alpha)(1-p_{00})p_{11}(1-p_{11})}{n((1-\alpha)(1-p_{00})+\alpha p_{11})(p_{00}+p_{11}-1)}\Bigg] + O\left(\frac{1}{n^2}\right).
\end{aligned} \tag{64}
$$

This concludes the proof of the bias and variance of the mixed estimator. Note that all terms of $O\left(\frac{1}{n^2}\right)$ are excluded.