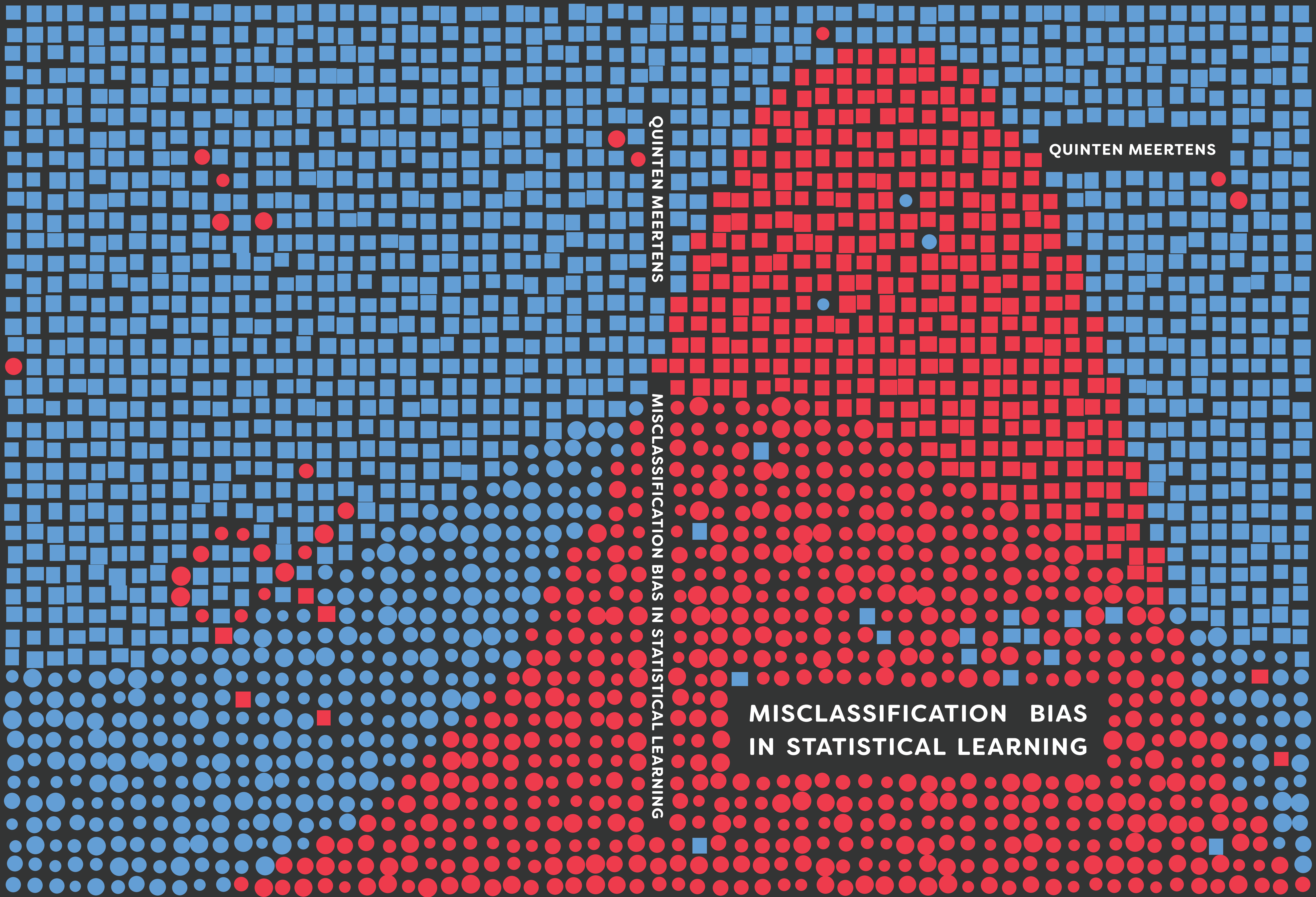
Misclassification bias in statistical learning

Meertens, Q.A.



QUINTEN MEERTENS

MISCLASSIFICATION BIAS IN STATISTICAL LEARNING

QUINTEN MEERTENS

MISCLASSIFICATION BIAS
IN STATISTICAL LEARNING

Misclassification Bias in Statistical Learning

Quinten Meertens

The author of this thesis was employed at Statistics Netherlands and used facilities at both the University of Amsterdam and Leiden University.

SIKS Dissertation Series No. 2021-10
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Typeset using LaTeX, figures generated using `Matplotlib` and `TikZ`
Cover design: Hans Westerbeek and Hendrik Zuidhoek, studiovlak.nl
Printed by: Gildeprint, Enschede, gildeprint.nl
ISBN: 978-94-6419-166-0

Misclassification Bias in Statistical Learning

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 28 april 2021, te 14.00 uur

door Quinten Alexander Meertens
geboren te Amsterdam

Promotiecommissie

Promotores:	prof. dr. C.G.H. Diks	Universiteit van Amsterdam
	prof. dr. H.J. van den Herik	Universiteit Leiden
Copromotores:	dr. F.W. Takes	Universiteit Leiden
Overige leden:	dr. K. Antonio	Universiteit van Amsterdam
	prof. dr. H.P. Boswijk	Universiteit van Amsterdam
	prof. dr. P.J.H. Daas	Technische Universiteit Eindhoven
	prof. dr. C.H. Hommes	Universiteit van Amsterdam
	prof. dr. E. Hüllermeier	Universität Paderborn
	prof. dr. J.J. Meulman	Universiteit Leiden

Faculteit Economie en Bedrijfskunde

Dedicated to Myrte, Nils, Olivier, and Thijmen

‘The ability to perceive or think differently is more important than the knowledge gained.’

David Bohm

PREFACE

After graduating in the summer of 2015, I was pondering whether to start a PhD or a job at Statistics Netherlands. Eventually, I chose the latter hoping to someday reignite my scientific spark. That day came sooner than I imagined. In 2016 I had the pleasure to meet my current supervisors Cees Diks, Jaap van den Herik, and Frank Takes. I recall that we connected on a scientific level as well as on a personal level and that we decided to start working together rather quickly.

Initially, our research focused on an applied problem: how can we accurately estimate cross-border Internet purchases within the European Union? It was a continuation of work that I took up earlier that year together with Arjan van Loon. When we achieved our first successful results in 2017 using statistical learning methods, Statistics Netherlands proposed to publish a news article on our work. As usual, the draft was reviewed by the Director-General, Tjark Tjin-a-Tsoi. His comments were: 'Amazing work. I only have a few remarks. First, we are suggesting impeccability, but I think that there must be at least some form of measurement error [...].'

As a consequence of this first remark, I soon discovered how misclassifications led to bias. I turned to my colleagues Arnout van Delden and Sander Scholtus. They had recently written a paper on the topic of classification errors together with Joep Burger, but they had not yet discovered the implications of their work to statistical learning. We decided to work on the topic together, for example by supervising Kevin Kloos, which resulted in Chapter 2 of this thesis. I truly enjoyed working closely with all four of you.

Next to having worked at Statistics Netherlands, I must say that I feel very privileged to have met so many inspiring colleagues at both CeNDEF (University of Amsterdam) and LIACS (Leiden University). Many of you directly or indirectly taught me about economics, econometrics and computer science, for which I am truly thankful.

Moreover, it has been a pleasure to work with and learn from you, Cees, Jaap and Frank. A personal expression of gratitude is included in the acknowledgements section, but I want to thank the three of you here as well. My PhD has been a great journey under your supervision and I believe your devotion to your students is exemplary. I look forward to our continued collaboration, in whatever form that may be.

Finally, I want to thank you, Myrte. A substantial part of this thesis was written during the COVID-19 pandemic. Expecting our third son, nourishing him for the past seven months, and having two other boys running around has been a huge challenge for both of us but for you in particular. I am so grateful for your love and your patience while I worked so hard to finish my thesis. We deserve our down time now, and I look forward to it.

Quinten Meertens
Almere, January 2021

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AB	AdaBoost
AUC	area under the receiver operator characteristic function
EU	European Union
FN	false negative
FP	false positive
GB	gradient boosting
kNN	k-nearest neighbours
LDA	linear discriminant analysis
LinSVC	linear support vector classification
LR	logistic regression
MNB	multinomial naive Bayes
MSE	mean squared (estimation) error
NACE	nomenclature statistique des activités économiques dans la Communauté Européenne
NSI	national statistical institute
OECD	Organisation for Economic Co-operation and Development
PS	problem statement
QDA	quadratic discriminant analysis
RBFSVC	support vector classification with radial basis function kernel
RF	random forest
RQ	research question
TN	true negative
TP	true positive

LIST OF SYMBOLS

C	column-normalised confusion matrix; calibration matrix
H	set of categories; set of classes
i	object in population I
I	indexed population
I_{test}	random sample for which true categories will be observed
K	number of categories
n	sample size of test set
n_{ij}, n_{i+}, n_{+j}	number of (in)correct predictions for objects in I_{test}
N	population size
N_{ij}, N_{i+}, N_{+j}	number of (in)correct predictions for the objects in I
p_{ij}	(mis)classification probabilities of classifiers
P	row-normalised confusion matrix
Q	inverse transposed row-normalised confusion matrix
$s_i, \hat{s}_i$	true and predicted category of object i
$v, \hat{v}$	number of objects in class of interest (within I)
$y_i, \hat{y}_i$	(estimated) numerical variable to aggregate
α	base rate
$\hat{\alpha}^*$	classify-and-count estimator for α
$\hat{\alpha}_a$	baseline estimator for α
$\hat{\alpha}_b$	subtracted-bias estimator for α
$\hat{\alpha}_c$	calibration estimator for α
$\hat{\alpha}_p$	misclassification estimator for α
γ, γ_β	AUC, softAUC
δ	prior probability shift
η, η_β	estimator of AUC, softAUC

CHAPTER 1

INTRODUCTION

1.1 Motivation

We live in a society that is driven by information. Currently, the most striking example is how governments deal with the COVID-19 pandemic. Lockdown rules and other restrictive measures are based predominantly on confirmed cases of coronavirus. The impact of these restrictive measures on our everyday lives is huge and therefore the swift availability of detailed and highly accurate information in this context is essential.

Confirmed cases of coronavirus are an example of statistical information about groups of people provided by the national government. Such information is referred to as *official statistics* and is often provided by national statistical institutes (NSIs). Over the past few years, NSIs have experienced an increase in the demand for official statistics, spanning across the following three dimensions: (1) statistics on new topics, including economic developments such as globalisation or the Internet economy, (2) more detailed statistics, both in space (e.g., small-area estimates) and time (e.g., at a higher frequency), and (3) more timely availability of statistics (Braaksma & Zeelenberg, 2015; De Broe et al., 2020). At the same time, NSIs not only experience budget cuts, but are also obliged to reduce the response burden on companies and citizens.

The consequence of these two conflicting developments (increased demand versus budget cuts) is that NSIs are in need for readily-available, detailed, and high-frequent data, which are often referred to as *big data*. Due to the high dimensionality of big data, new statistical methods have to be developed as well (Hastie, Tibshirani & Friedman, 2009). These methods are often referred to as methods in data science, machine learning or artificial intelligence. We prefer the term *statistical learning*, following Hastie et al. (2009). Although big data and statistical learning have been popular topics in the quantitative sciences for at least two decades (the first edition of *The Elements of Statistical Learning* was published in 2002), NSIs have started embracing them only a few years ago. Briefly put, the main obstacles have been the quality of the data and the quality of the methods (Struijs, Braaksma & Daas, 2014).

This thesis focuses on understanding and improving the quality of the methods. We will concentrate on a specific impediment called misclassification bias. This is a type of statistical bias that occurs when imperfect classifications are subsequently counted or otherwise aggregated. The bias might be large, even for highly accurate statistical learning methods. Therefore, we believe that this thesis contributes to a more thorough evaluation of statistical learning methods, enabling more reliable use of these methods in official statistics in particular and the quantitative sciences in general.

Below we provide a brief introduction to official statistics, statistical learning and misclassification bias. In Section 1.2, we describe how NSIs nowadays embrace model-based statistics besides design-based statistics. In Section 1.3, we discuss statistical learning in official statistics. In Section 1.4, we introduce misclassification bias and indicate an open problem in the academic literature. We then formulate the main problem statement and the three research questions in Section 1.5. The research methodology is given in Section 1.6. The four contributions of the thesis are listed in Section 1.7. Finally, the structure of the thesis is outlined in Section 1.8.

1.2 Official statistics

NSIs all over the world produce statistical information on social and economic issues, including economic growth, safety and healthcare. Such information is referred to as official statistics. The aims and scope of NSIs are based on fundamental principles that are agreed upon within international and intergovernmental organisations, including the United Nations (Statistical Commission of the United Nations, 2014) and the OECD (OECD, 2011). Within the European Union, the fundamental principles are established in the Regulation on European Statistics (European Commission, 2009) and in the European Statistics Code of Practice (Eurostat, 2017). The fundamental principles that relate to the output quality of official statistics (including statistical accuracy) were a reason for NSIs to restrict themselves to design-based statistics, as opposed to model-based statistics (Braaksma & Zeelenberg, 2015).

However, three interrelated developments have initiated and accomplished a paradigm shift within official statistics over the last two decades (Buelens, Boonstra, van den Brakel & Daas, 2012; De Broe et al., 2020). The first development is the *rise of big data*. These data are not suitable for classical design-based statistics due to the issue of selectivity (Daas, Puts, Buelens & Van den Hurk, 2015; Van den Brakel & Bethlehem, 2008). The second development is the *sophistication of statistical learning methods*, including both statistical models and learning algorithms, which currently are able to deal with the high dimensionality encountered in big data (Hastie et al., 2009). The third development, the significance of which is often underestimated, is the *implementation of learning algorithms into easy-to-use open source software*. A noteworthy example is the scikit-learn module in Python (Pedregosa et al., 2011). Nowadays, one can employ the most complex learning algorithms within a few lines of code, whereas just over a decade ago the complexity of such models hampered their adoption into the production process of official statistics (Van den Brakel & Bethlehem, 2008).

The consequence of these three developments is that NSIs are now embracing both model-based statistics and algorithm-based statistics. The most profound example is the production of the consumer price index. A few years ago, experimentation with web scraping and scanner data was initiated at Statistics Netherlands (Chessa, 2016) and as of 2020 no other sources of information are used anymore[1]. It should be stressed that NSIs still comply to the same fundamental principles of official statistics, also for the production of model-based or algorithm-based statistics, albeit by adopting additional guidelines (Buelens, de Wolf & Zeelenberg, 2016). Therefore, it might be argued that the paradigm in official statistics is not shifting but rather being updated (cf. De Broe et al., 2020).

[1] See `https://www.cbs.nl/en-gb/corporate/2020/02/manual-retail-price-observations-discontinued`.

Nonetheless, the adoption of models and algorithms in the field of official statistics generates a wide variety of methodological challenges concerning the quality of the resulting statistical output. These methodological challenges are discussed next.

1.3 Statistical learning in official statistics

The rise of big data has sparked the adoption of statistical learning methods at NSIs worldwide. Recently, Beck, Dumpert and Feuerhake (2018) reported 136 ongoing machine learning projects at NSIs in 25 countries, most of which (78) aim at producing new or improved statistical output using classification algorithms. Despite the fact that NSIs seem to be focused on creating new output, the methodological challenges of applying statistical learning methods to big data are widely recognised (see, e.g., Kitchin, 2015 and MacFeely, 2016)

From the viewpoint of Statistics Netherlands, De Broe et al. (2020) address the most significant challenges encompassing the use of statistical learning in the production of official statistics. Many of the methodological challenges they present are due to biases encountered in big data sources, which have been described extensively in the scientific literature recently (Baeza-Yates, 2018; Mehrabi, Morstatter, Saxena, Lerman & Galstyan, 2019). Moreover, De Broe et al. (2020) summarise four challenges encompassing big data *methods* (cf. statistical learning methods) in official statistics. The four challenges are dealing with (1) noise, (2) selectivity, (3) spurious correlations, and (4) concept drift. All four of these have impact on the classification accuracy of the statistical learning methods used. In this thesis, we will devote specific attention to concept drift. It is defined as a change in the joint distribution of the dependent and independent variables (Gama, Žliobaitė, Bifet, Pechenizkiy & Bouchachia, 2014; Webb, Hyde, Cao, Nguyen & Petitjean, 2016). As official statistics often describe non-stationary stochastic processes, like economic growth, concept drift always occurs, even if the data contain no errors or biases.

Subsequently, De Broe et al. (2020) provide a general quality framework for the use of big data and statistical learning in the production of official statistics, complementing the guidelines by Buelens et al. (2016) mentioned above. Within that general quality framework, this thesis focuses on *improving the accuracy* (as part of the output quality) of what we will refer to as *classifier-based statistics*, i.e., aggregate statistics that are based on categorical variables predicted by statistical learning methods.

There are two main reasons for focusing on the accuracy of classifier-based statistics. First, improving the accuracy of classifier-based statistics also contributes to other scientific disciplines that are interested in aggregate statistics, including epidemiology, remote sensing and political science. In fact, the results that we have obtained are not restricted to big data sources, but apply to classifier-based statistics regardless of which type of data is used. Second, the underlying problem that causes inaccuracy of classifier-based statistics, namely misclassification, has always been a neglected problem in categorical data analysis (Schwartz, 1985). Even today, retrieving aggregate statistics from learning algorithms is mistakenly believed to be a trivial task, leading to severe statistical bias (González, Castaño, Chawla & del Coz, 2017).

In Section 1.4, we show how misclassification may lead to statistical bias. We then indicate an open problem in the academic literature on misclassification bias in statistical learning. It leads to the formulation of the problem statement of this thesis in Section 1.5.

1.4 Misclassification bias in statistical learning

Misclassification bias is defined as the statistical bias of an estimator of an aggregate statistic, which results from errors in the classifications on the level of individual objects (Czaplewski, 1992). An illustrative example is provided in the box titled “The Election Prediction Example”. The example shows that misclassifications do *not* cancel out by aggregation, but yield biased estimates of aggregate statistics. This is in contrast to what is sometimes claimed (cf. O’Connor, Balasubramanyan, Routledge and Smith, 2010, p. 125). Estimating aggregate statistics using statistical learning methods is referred to as the *quantification task for machine learning* (Forman, 2008):

> *“Given a limited training set with class labels, induce a quantifier that takes an unlabeled test set as input and returns its best estimate of the number of cases in each class.”* (Forman, 2008, p. 167)

The quantification task for machine learning (in brief, *quantification learning*) was first described by Forman (2005). His key observation was that accurate classification is not necessary for accurate quantification. Indeed, it is true to some extent, as depicted in Fig. 1.1, adapted from Scholtus and Van Delden (2020). The figure shows that increasing classification accuracy might increase misclassification bias, but will also often reduce standard deviation.

The Election Prediction Example. (*adapted from Meertens, Van Delden, Scholtus and Takes, 2019*).

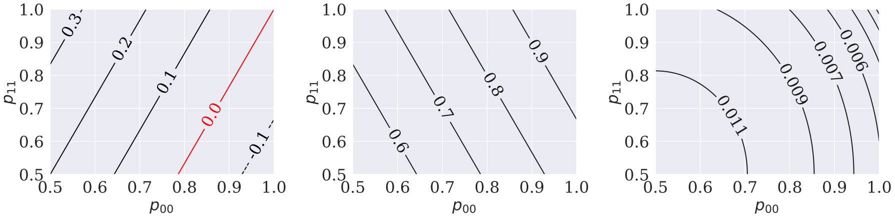
Assume that we are interested in predicting the outcome of the popular vote of the US election between two candidates, named A and B. We assume that 140 million (M) people will vote, of which we assume 66M will vote for candidate A and 74M for candidate B. Clearly, candidate B will become the winner of the popular vote.

To predict the election outcome, we consider opinion polling based on Twitter data (Jaidka, Ahmed, Skoric & Hilbert, 2018; O'Connor et al., 2010). We assume that we have trained a classifier predicting the candidate preference for each voter based on the voter's tweets (for the purpose of this example, it is assumed that all voters are active on Twitter). We write p_{AA} for the probability that the classifier correctly predicts the political preference of a voter that will vote for candidate A. We define p_{BB} similarly. We use v_A for the *actual* number of voters for candidate A and $\widehat{v}_A$ for the classifier's *predicted* number of voters for candidate A. We use similar notation for candidate B.

Taking the accuracy scores from the sentiment analysis performed by Jaidka et al. (2018), we obtain misclassification probabilities $p_{AA} = 0.93$ and $p_{BB} = 0.87$. Based on such a classifier, the expected numbers of predicted votes for candidates A and B are given by

$$\begin{cases} \mathbb{E}(\widehat{v}_A) = p_{AA} \cdot v_A + (1 - p_{BB}) \cdot v_B, \\ \mathbb{E}(\widehat{v}_B) = (1 - p_{AA}) \cdot v_A + p_{BB} \cdot v_B, \end{cases} \tag{1.1}$$

which yields $\mathbb{E}(\widehat{v}_A) = 71$M and $\mathbb{E}(\widehat{v}_B) = 69$M (with negligible standard deviation). It means that the algorithm would predict the **wrong winner** of the popular vote, despite being a rather accurate predictor of the political preference of individual voters.



(A) Misclassification bias (B) Classification accuracy (C) St. dev. ($N = 2{,}000$)

FIG. 1.1: Contour lines showing the difference between (A) quantification performance and (B) classification performance for base rate $\alpha = 0.3$ and varying classification probabilities p_{00} and p_{11}. The quantification performance is maximised at the red line (no bias). The key observation is that the directions of the contour lines are opposite (and will be perpendicular if $\alpha = 0.5$). Panel (C) depicts the standard deviation of classifier-based statistics. Observe how reducing the classification accuracy might also reduce misclassification bias, but will always increase standard deviation. *Panels (A) and (C) adapted from Scholtus and Van Delden (2020).*

Moreover, we may observe that producing official statistics based on statistical learning methods is the exact same task as the quantification task for machine learning. Therefore, methods for quantification learning are immediate candidates for reducing misclassification bias in classifier-based statistics. A recent taxonomy of quantification learning methods is provided by González et al. (2017). They distinguish between three categories of methods: those that (1) correct classifier-based statistics, (2) adjust classifiers, and (3) match class distributions. Many methods in these three categories have been proposed and they have been evaluated empirically, but, as González et al. (2017) stress in their conclusions,

> *"First, more solid theoretical analyses are needed to better understand not only the behavior of these algorithms but also the learning problem in general."* (González et al., 2017, p. 74:37),

so theoretical evaluations are missing.

In this thesis we will provide theoretical analyses for the first category of methods. Our starting point is the statistical literature on misclassification bias in categorical data analysis, dating back to at least the seminal work by Bross (1954). The literature is rather comprehensive, as the overlapping period is almost 70 years. Therefore, we will rely on the relatively recent overview provided by Buonaccorsi (2010). There, the misclassifications are attributed to measurement errors instead of prediction errors. Nonetheless, the mathematical descriptions of misclassification bias in (a) categorical data analysis and (b) quantification learning are identical. The mathematical description used is captured by three ingredients: (i) a statistical model for the true values, (ii) a measurement error model, and (iii) extra data, information or assumptions needed to correct for measurement error (Buonaccorsi, 2010, p. 4). In this thesis, we take the three ingredients as our basis. We substantiate the three ingredients by rather precise descriptions.

(i) As a first step, we focus on binary classification problems. With this focus, the true values (categories) are *independent and identically distributed (iid)* random variables with a Bernoulli(α)-distribution. The objective is to estimate α, which is called the *base rate*.

(ii) The measurement error model we use is the *classical measurement error model* (as opposed to the Berkson error model, see Berkson, 1950), which entails that we make assumptions on the distribution of observed values given the true values (and not vice versa) (Buonaccorsi, 2010, p. 6). Here, we follow the convention in official statistics (see Van Delden, Scholtus and Burger, 2016) and not that in remote sensing (cf. Czaplewski, 1992). Our motivation is that an observation (i.e., prediction) is causally determined, with error, by the underlying true category. For example, consider epidemiology, where a disease causally determines the symptoms and hence the test result, not vice versa.

(iii) Out of the five typical examples of extra data presented by Buonaccorsi (2010), we will take our validation data. In the context of statistical learning, such data are called *test data*. In supervised learning problems, test data are already available to estimate the model's out-of-sample performance. Therefore, choosing test data as extra data to correct for measurement error is a natural choice.

Finally, the open problem in the statistical literature on misclassification bias is that *for finite populations* no theoretical comparison between correction methods is available. The focus seems to have been on asymptotic results only (Kuha & Skinner, 1997). Although the populations typically considered in official statistics are quite large, the samples used as test data are often small. When correcting misclassification bias, the size of the test data (and not that of the population) determines the accuracy of the estimate of the base rate α. We will use the mean squared (estimation) error (MSE) to measure estimation accuracy, following both the convention in official statistics (Buelens et al., 2016) and the recommendation in quantification learning (Sebastiani, 2020). We are now able to formulate the problem statement and research questions of this thesis.

1.5 Problem statement and three research questions

As we have argued so far, NSIs increasingly embrace statistical learning methods to produce official statistics, but the output quality is hampered by misclassification bias. This observation gives us the opportunity to formulate the following problem statement.

> **Problem statement (PS)**: *In what way can we reduce misclassification bias in statistical learning so that we obtain more accurate classifier-based statistics?*

By using the MSE to measure the accuracy of classifier-based statistics, we take the bias-variance trade-off into account that arises when reducing misclassification bias. We will derive two theoretical research questions and one applied research question from the PS.

First research question. In Section 1.4, we remarked that for finite populations no theoretical comparison between the correction methods for misclassification bias is available. Therefore, the first research question reads as follows.

> **Research question 1 (RQ1)**: *Which estimator of the base rate, in particular when dealing with concept drift, has the smallest MSE in finite populations?*

Second research question. The book by Buonaccorsi (2010) explicitly excludes Bayesian methods for categorical data analysis. Still, there are at least two reasons to consider Bayesian methods. The first reason is that Bayesian methods can solve the identification problem that occurs if the misclassification probabilities are not known exactly (Gaba & Winkler, 1992). The second reason is that Bayesian methods allow the use of prior information in a natural way. More specifically, the prior information can be used *"to derive identification regions for any real functional* *of the distribution of interest"* (Molinari, 2008). In many applications in statistics (e.g., in epidemiology, see Goldstein et al., 2016 and Gustafson, 2004) these identification regions are not yet leveraged to improve the accuracy of classifier-based statistics. Thus, our second research question reads as follows.

> **Research question 2 (RQ2)**: *How can we leverage identification regions of misclassification probabilities in order to reduce the MSE of classifier-based statistics even further?*

Third research question. The answers to the first two research questions provide a theoretical answer to the problem statement. To complement it, we will consider a specific application in official statistics, namely the problem of estimating cross-border Internet purchases within the European Union (EU). That estimation problem has been a challenge in official statistics for many years and still is due to a lack of sufficiently accurate data. We will investigate how statistical learning can be used to improve the accuracy of existing estimates of cross-border Internet purchases by answering the following research question.

> **Research question 3 (RQ3)**: *To what extent can statistical learning be used to improve the accuracy of estimates of cross-border Internet purchases within the EU?*

1.6 Research methodology

The research methodology that we will follow to answer each of the three research questions is given below. We also provide an overview in Table 1.1.

First, we will answer RQ1 by means of literature review, theoretical derivations, and numerical analyses. We will do so in two steps, namely under two different assumptions. The first assumption (A1) is that the test data are a random sample from the population. The assumption corresponds to the *double sampling scheme* introduced by Tenenbein (1970). The second assumption (A2) is that the class distribution might differ between the test data and the unlabelled data, but that the misclassification probabilities are the same. Assumption A2 corresponds to a specific type of concept drift referred to as *prior probability shift* (Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla & Herrera, 2012). Assumption A1 can be viewed as assumption A2 with the additional restriction that the prior probability shift

Table 1.1: Overview of the research methodologies used to answers each of the three research questions.

Research methodology	*RQ1*	*RQ2*	*RQ3*
Literature review	✓	✓	✓
Theoretical derivations	✓	✓	
Numerical analyses	✓		
Simulation study		✓	
Experimental study			✓

is zero. We will first answer RQ1 *under assumption A1* by means of literature review and theoretical derivations. Then, we will answer RQ1 *under assumption A2* by reviewing the literature, adapting the theoretical derivations that resulted from answering RQ1 under assumption A1, and examining the adapted formulas numerically.

Subsequently, we will answer RQ2 by a review of the literature, theoretical derivations, and a simulation study. Our starting point will be the overview of Bayesian methods for categorical data analysis provided by Agresti and Hitchcock (2005). We will then extend the analytical expressions for the posterior distribution (of misclassification probabilities) under conjugate priors to our setting of misclassification bias. Based on these expressions, we will provide a simulation study to examine the MSE of classifier-based statistics when leveraging identification regions as prior restrictions.

Finally, we will answer RQ3 by a literature review and an extensive experimental study. We will discover that existing methods to estimate cross-border Internet purchases within the EU are based on demand-side data and that the use of these data results in an underestimation. Hence, we will propose a new methodology to estimate cross-border Internet purchases based on (1) supply-side data and (2) state-of-the-art methods from both computer science and statistical learning. We will then implement the proposed methodology for the case of the Netherlands and compare the resulting estimates with existing estimates. The comparison will provide the empirical evidence to answer RQ3.

1.7 Four contributions

Below, we list the four main contributions of this thesis.

- *Contribution 1.* We prove which estimation method minimises the MSE of classifier-based statistics when correcting misclassification bias in finite populations.

  We provide new theoretical derivations for the MSE of two popular corrections methods (the misclassification estimator and the calibration estimator, see Kuha and Skinner, 1997). We do so under the two assumptions A1 and A2 (see Section 1.6). Under assumption A1, we provide a conclusive theoretical result showing that the MSE of the calibration estimator is always below that of the misclassification estimator. The result will be referred to as Contribution 1(a). Under assumption A2, we show that the estimator that minimises the MSE might be either the misclassification estimator or the calibration estimator, depending on the data and the model used. This result will be referred to as Contribution 1(b).

- *Contribution 2.* We reduce the MSE of the misclassification estimator even further by leveraging identification regions as prior restrictions.

  A Bayesian framework for reducing misclassification bias is considered. Within that framework we propose prior restrictions on misclassification probabilities based on identification regions (Molinari, 2008). A simulation study shows that the prior restrictions reduce the MSE even further, in particular when dealing with small test data sets. Moreover, our construction guarantees that impermissible estimates (e.g., negative counts) are prevented.

- *Contribution 3.* We propose an internationally consistent and comparable method to estimate cross-border Internet purchases within the EU, which is more accurate than existing methods.

  We identified three supply-side data sources that capture information on cross-border Internet purchases within the EU. We applied probabilistic record linkage (Fellegi & Sunter, 1969) through approximate string matching (Cohen, Ravikumar & Fienberg, 2003) and locality-sensitive hashing (Bawa, Condie & Ganesan, 2005; Broder, 1997) to combine the three data sources in an accurate and timely manner. Moreover, we employed statistical learning methods to identify webshops from the combined data and implemented methods to correct for the resulting misclassification bias. Our method is consistent and comparable across EU member states. We applied our method to the Netherlands for the year 2016 and found an estimate (with a standard deviation of only 8 percent) that is 6 times as high as estimates produced by existing methods.

- *Contribution 4.* We show theoretically that smoothing the model-performance metric AUC might reduce its MSE, contradicting empirical evidence.

  The area under the receiver operating characteristic curve (AUC) is a performance metric used to evaluate algorithms that *rank* instead of *classify* objects according to a dichotomous variable. We consider a smoothed variant of the AUC proposed by Yan, Dodier, Mozer and Wolniewicz (2003). This variant of the AUC is claimed to be outperformed by the standard AUC as a model selector for algorithms that rank objects (Vanderlooy & Hüllermeier, 2008). The claim is supported by empirical evidence only. We therefore investigated the theoretical properties of the standard AUC and those of the smoothed variant of the AUC. Based on our investigations, we present preliminary theoretical derivations that seem to contradict the claim.

1.8 Outline of the thesis

In Chapter 1, we connected official statistics, quantification learning and classical categorical data analysis. We introduced the problem of misclassification bias in statistical learning and we formulated the problem statement and three research questions of the thesis. We then highlighted our four contributions. The remainder of this thesis is organised as follows.

- **Chapter 2** will answer RQ1 under assumption A1. The main result of the chapter is Contribution 1(a). The content is based on the paper titled “Comparing correction methods to reduce misclassification bias” by K. Kloos, Q.A. Meertens, S. Scholtus and J.D. Karch (2020), published in the proceedings (peer-reviewed) of the 32nd Benelux Conference on Artificial Intelligence and Machine Learning (BNAIC), edited by L. Cao, W.A. Kosters and

J. Lijffijt, pages 103–129. Moreover, the paper was selected for the postproceedings to be published as part of Springer's Communications in Computer and Information Science (CCIS) series. (The paper was also nominated for the Best Paper Award of the conference.)

- **Chapter 3** will answer RQ1 under assumption A2. The result corresponds to Contribution 1(b). The chapter's content is identical to the manuscript titled "Improving the output quality of official statistics based on machine learning algorithms" by Q.A. Meertens, C.G.H. Diks, H.J. van den Herik and F.W. Takes, submitted to the Journal of Official Statistics in December 2020.

- **Chapter 4** will answer RQ2 and hence result in Contribution 2. The content is the same as the paper titled "A Bayesian approach for accurate classification-based aggregates" by Q.A. Meertens, C.G.H. Diks, H.J. van den Herik and F.W. Takes (2019), published in the proceedings (peer-reviewed) of the 19th SIAM International Conference on Data Mining (SDM), edited by T.Y. Berger-Wolf and N.V. Chawla, pages 306–314.

- **Chapter 5** will provide an answer to RQ3. The answer corresponds to Contribution 3. The content of the chapter is identical to the paper titled "A data-driven supply-side approach for estimating cross-border Internet purchases within the European Union" by Q.A. Meertens, C.G.H. Diks, H.J. van den Herik and F.W. Takes (2020), published in the Journal of the Royal Statistical Society, Series A (Statistics in Society), **183**(1), pages 61–90.

- **Chapter 6** will touch upon an alternative to classification, namely ranking. The chapter positions a theoretical open problem on model selection of rankers, as opposed to classifiers. It has Contribution 4 as the main result. The content of the chapter is ongoing work that is not yet submitted, but it is included in this thesis as a discussion paper.

- **Chapter 7** will provide the conclusions of this thesis in three parts. We will (1) answer the three research questions, (2) answer the problem statement, and (3) discuss directions for future research.

CHAPTER 2

DOUBLE SAMPLING SCHEME

2.1 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 (Curier et al., 2018). 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 (Scholtus & Van Delden, 2020; Schwartz, 1985). 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* (Czaplewski, 1992).

Misclassification bias occurs in a broad range of applications, including official statistics (Meertens, Diks, Van den Herik & Takes, 2020), land cover mapping (Löw, Knöfel & Conrad, 2015), political science (Hopkins & King, 2010; Wiedemann, 2019), and epidemiology (Greenland, 2014). The objective in each of these applications is to minimise a loss function at the level of aggregated predictions, in contrast to minimising 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 González et al. (2017) 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 data sets are needed to optimise 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 (González et al., 2017). 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 Kuha and Skinner (1997) for a technical discussion and Buonaccorsi (2010) 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 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 data sets. 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 data sets. 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 organised as follows. First, in Section 2.2, the five estimators are formally introduced and the mathematical expressions for their MSEs are presented. The derivations are provided in Appendix 2.A. Then, in Section 2.3, 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 smallest MSE. Finally, in Section 2.4, we draw our main conclusion and discuss directions for future research.

2.2 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 2.1, the objects are houses in the Netherlands and the two classes are whether or not the house has solar panels on the roof (Curier et al., 2018). 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 2.2.2, 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* (Czaplewski, 1992). In this section, five estimators for the base rate parameter α are formally introduced, many of which have been proposed decades ago, see Kuha and Skinner (1997) for an extensive discussion. We summarise 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 Van Delden et al. (2016), 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* P is then defined as follows:

$$P = \begin{pmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{pmatrix}. \tag{2.1}$$

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 2.1. 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

Table 2.1: Contingency tables for test set (left) and target population (right).

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

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

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 2.1, 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 **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.

2.2.1 Baseline estimator - random sample

The baseline estimator of α is the proportion of data points in the test data set 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 of α, i.e.:

$$B\left[\hat{\alpha}_a\right] = 0. \tag{2.2}$$

Since we have assumed that the size n of the test data set 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

$$MSE\left[\hat{\alpha}_a\right] = V\left[\hat{\alpha}_a\right] = \frac{\alpha(1-\alpha)}{n}. \tag{2.3}$$

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

2.2.2 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 of α, which we will denote by $\hat{\alpha}^*$, is referred to as the 'classify-and-count' estimator, see González et al. (2017). In general, the classify-and-count estimator is (strongly) biased, and has almost zero variance. More specifically,

$$\mathbb{E}\left[\hat{\alpha}^*\right] = \alpha p_{11} + (1-\alpha)(1-p_{00}), \tag{2.4}$$

and hence

$$B\left[\hat{\alpha}^*\right] = \alpha(p_{11}-1) + (1-\alpha)(1-p_{00}), \tag{2.5}$$

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 by Scholtus and Van Delden (2020). The variance of the classify-and-count estimator is derived by Burger, Van Delden and Scholtus (2015) and equals

$$V\left[\hat{\alpha}^*\right] = \frac{\alpha p_{11}(1-p_{11}) + (1-\alpha)p_{00}(1-p_{00})}{N}. \tag{2.6}$$

If the population size N is large, the variance of $\hat{\alpha}^*$ is small. In some literature, this small variance is misinterpreted as high accuracy, by claiming intuitively that the large size of the data set implies that the noise cancels out (cf. O'Connor et al., 2010). 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

$$MSE\left[\hat{\alpha}^*\right] = \left[\alpha(p_{11}-1) + (1-\alpha)(1-p_{00})\right]^2 + O\left(\frac{1}{N}\right). \tag{2.7}$$

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., Strichartz (2000, p. 147). Observe how, in general, the MSE does not converge to 0 as N tends to ∞.

2.2.3 Subtracting estimated bias

Knowing that the classify-and-count estimator $\hat{\alpha}^*$ is biased (see (2.5)), we may attempt to estimate that bias and subtract it from $\hat{\alpha}^*$. As briefly mentioned by Scholtus and Van Delden (2020), we may estimate that bias by the plug-in estimator, that is, we substitute the unknown quantities in expression (2.5) by their estimates. More precisely, the bias is estimated as

$$\widehat{B}\left[\hat{\alpha}^*\right] = \hat{\alpha}^*(\hat{p}_{00} + \hat{p}_{11} - 2) + (1 - \hat{p}_{00}), \tag{2.8}$$

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

$$\hat{\alpha}_b = \hat{\alpha}^* - \widehat{B}\left[\hat{\alpha}^*\right] = \hat{\alpha}^*(3 - \hat{p}_{00} - \hat{p}_{11}) - (1 - \hat{p}_{00}). \tag{2.9}$$

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

$$B\left[\hat{\alpha}_b\right] = (1 - p_{00})(2 - p_{00} - p_{11}) - \alpha(p_{00} + p_{11} - 2)^2. \tag{2.10}$$

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

$$\begin{aligned} V\left[\hat{\alpha}_b\right] = {} & \frac{\left[\alpha(p_{00} + p_{11} - 1) - p_{00}\right]^2 p_{00}(1 - p_{00})}{n(1 - \alpha)}\left(1 + \frac{\alpha}{n(1 - \alpha)}\right) \\ & + \frac{\left[\alpha(p_{00} + p_{11} - 1) + (1 - p_{00})\right]^2 p_{11}(1 - p_{11})}{n\alpha}\left(1 + \frac{1 - \alpha}{n\alpha}\right) \\ & + O\left(\max\left[\frac{1}{n^3}, \frac{1}{N}\right]\right). \end{aligned} \tag{2.11}$$

Proof. See Appendix 2.A. □

In particular, Theorem 2.1 implies that $B\left[\hat{\alpha}_b\right] = (2 - p_{00} - p_{11})B\left[\hat{\alpha}^*\right]$, compare Equations (2.10) and (2.5). Hence, $\left|B\left[\hat{\alpha}_b\right]\right| \leq \left|B\left[\hat{\alpha}^*\right]\right|$, because $1 < p_{00} + p_{11} < 2$.

2.2.4 Misclassification probabilities

Let P be the row-normalised confusion matrix of the machine learning algorithm that we have trained, as defined in (2.1). 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 $\boldsymbol{\alpha}^*$). Under the assumption that the probabilities p_{ij} are identical for each data point, we obtain the expression $\mathbb{E}[\boldsymbol{\alpha}^*] = P^T\boldsymbol{\alpha}$. If the true values of all entries p_{ij} of P were known and if $p_{00}+p_{11} \neq 1$, then $\boldsymbol{\alpha}_p = (P^T)^{-1}\boldsymbol{\alpha}^*$ would be an unbiased estimator of α. Using the plug-in estimator $\hat{P}$ for P, estimated on the test set, the following estimator of α is obtained:

$$\hat{\alpha}_p = \frac{\hat{\alpha}^* + \hat{p}_{00} - 1}{\hat{p}_{00} + \hat{p}_{11} - 1}. \tag{2.12}$$

It is known that the estimator $\hat{\alpha}_p$ is consistent (asymptotically unbiased) for α, see Buonaccorsi (2010). Grassia and Sundberg (1982) analysed the variance of this estimator of an arbitrary number of classes. For the binary case, a simple analytic expression for the bias and variance of $\hat{\alpha}_p$ for finite data sets has not been given, as far as we know. Therefore, we have derived the bias and variance for finite data sets, yielding the following result.

Theorem 2.2. *The bias of $\hat{\alpha}_p$ as estimator of α is given by*

$$B\left[\hat{\alpha}_p\right] = \frac{p_{00} - p_{11}}{n(p_{00} + p_{11} - 1)} + O\left(\frac{1}{n^2}\right). \tag{2.13}$$

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

$$V[\hat{\alpha}_p] = \frac{(1-\alpha)p_{00}(1-p_{00})\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + \alpha p_{11}(1-p_{11})\left[1 + \frac{1-\alpha}{n\alpha}\right]}{n(p_{00}+p_{11}-1)^2} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right). \tag{2.14}$$

Proof. See Appendix 2.A. □

2.2.5 Calibration probabilities

Let C be the column-normalised 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 (Kuha & Skinner, 1997). The first element of the vector $C\boldsymbol{\alpha}^*$ is an unbiased estimator of $\boldsymbol{\alpha}$, if C is known.

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

$$\hat{\alpha}_c = \hat{\alpha}^* \frac{n_{11}}{n_{+1}} + (1 - \hat{\alpha}^*) \frac{n_{10}}{n_{+0}}, \tag{2.15}$$

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 of α (Buonaccorsi, 2010). Under the assumptions we have made in this paper, it can be shown that $\hat{\alpha}_c$ is in fact an unbiased estimator of 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 summarised in the following theorem.

Theorem 2.3. *The calibration estimator $\hat{\alpha}_c$ is an unbiased estimator of α:*

$$B\left[\hat{\alpha}_c\right] = 0. \tag{2.16}$$

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

$$\begin{aligned}
V(\hat{\alpha}_c) = &\left[\frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n} + \frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n^2}\right] \\
&\times \left[\frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}} \left(1 - \frac{\alpha p_{11}}{(1-\alpha)(1-p_{00}) + \alpha p_{11}}\right)\right] \\
&+ \left[\frac{(1-\alpha)p_{00} + \alpha(1-p_{11})}{n} + \frac{(1-\alpha)(1-p_{00}) + \alpha p_{11}}{n^2}\right] \\
&\times \left[\frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})} \left(1 - \frac{(1-\alpha)p_{00}}{(1-\alpha)p_{00} + \alpha(1-p_{11})}\right)\right] \\
&+ O\left(\max\left[\frac{1}{n^3}, \frac{1}{Nn}\right]\right).
\end{aligned} \tag{2.17}$$

Proof. See Appendix 2.A. □

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 MSE of the five estimators, both theoretically as well as by means of simulation studies.

2.3 Results

The aim of this section is to derive empirically which of the five estimators of α that we presented in Section 2.2 has the smallest 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 visualisations 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 box plots. 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 smallest MSE.

2.3.1 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 data set, that is, $\alpha = 0.5$, with a small test data set of size $n = 1000$, a large population data set $N = 3 \times 10^5$ and a rather poor classifier having classification probabilities $p_{00} = 0.6$ and $p_{11} = 0.7$. We choose $p_{00} \neq p_{11}$ deliberately, as otherwise the classify-and-count estimator $\hat{\alpha}^*$ would be unbiased, i.e., (p_{00}, p_{11}) would be on the line between $(1 - \alpha, \alpha)$ and $(1, 1)$, see also expression (2.5).

Table 2.2 summarises the bias, variance and MSE, computed using the analytic approximations presented in Section 2.2. The classify-and-count estimator is highly biased and therefore it has a large MSE, despite having the smallest variance of all estimators. The MSE 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-normalised confusion matrix P (that is, the misclassification probabilities) for

Table 2.2: Comparison of the bias, variance and MSE of each of the five estimators of the base rate α when encountering no class imbalance.[†]

Estimator	*Symbol*	*Bias* $(\times 10^{-2})$	*Variance* $(\times 10^{-4})$	*MSE* $(\times 10^{-4})$
Baseline	$\hat{\alpha}_a$	0.000	2.500	2.500
Classify-and-count	$\hat{\alpha}^*$	5.000	0.000	25.000
Subtracted-bias	$\hat{\alpha}_b$	3.500	2.244	14.494
Misclassification	$\hat{\alpha}_p$	-0.033	25.025	25.026
Calibration	$\hat{\alpha}_c$	0.000	2.275	2.275

[†] Simulation results for $\alpha = 0.5$, $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$.

values of p_{00} and p_{11} close to $p_{00} + p_{11} = 1$ significantly increases the variance of the estimator, leading to the largest MSE of all estimators considered. Finally, the calibration estimator $\hat{\alpha}_c$ is unbiased and has the smallest variance among the estimators that make use of the test data set. In particular, note that the variance is also smaller than that of the baseline estimator. In this example, the estimator based on the calibration probabilities has the smallest MSE, and it is the only estimator with a smaller MSE than the baseline estimator $\hat{\alpha}_a$.

To gain insight in the sampling distribution of the estimators, in addition to the metrics presented in Table 2.2, we simulated a large number $R = 10{,}000$ of confusion matrices for data sets 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 and a random draw from a $Bin(N - N_{+1}, p_{00})$-distribution to obtain N_{11} and N_{00} (respectively). 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 $R = 10{,}000$ times gave rise to the sampling distributions of the five estimators as presented in Fig. 2.1. It nicely visualises the bias and variance of the five estimators, supporting the results in Table 2.2. 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.

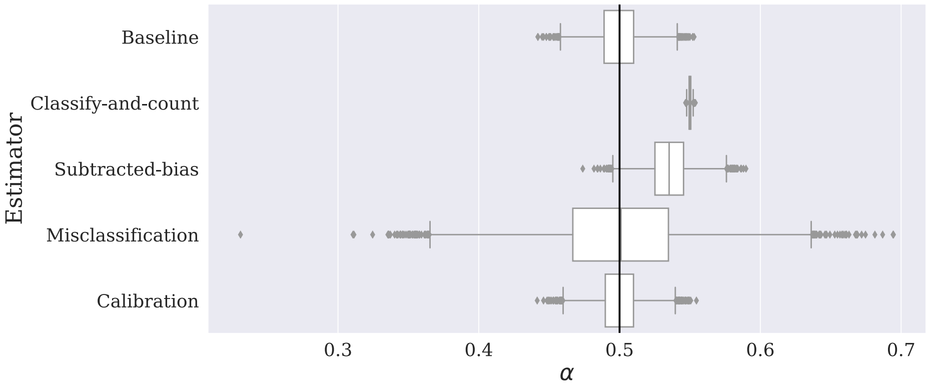


FIG. 2.1: The box plots show the sampling distribution of the estimators for α, where $\alpha = 0.5$, $p_{00} = 0.6$, $p_{11} = 0.7$, $n = 1000$ and $N = 3 \times 10^5$. The true value of α is highlighted by a vertical line.

In the second simulation study, we consider a highly imbalanced data set, namely $\alpha = 0.98$. We again assume that the available test data set has size $n = 1000$, but we assume a classifier having classification probabilities $p_{00} = 0.94$ and $p_{11} = 0.97$. Table 2.3 summarises the bias, variance and MSE of each of the estimators and Fig. 2.2 shows the sampling distributions of each of the estimators. 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 data set is highly imbalanced. Its MSE is slightly larger than the MSE of the method with calibration probabilities and much smaller 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.

2.3.2 Finding the optimal estimator

The aim of this subsection is to find the optimal estimator, i.e., the estimator with the smallest MSE, 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

Table 2.3: Comparison of the bias, variance and MSE of each of the five estimators of the base rate α when encountering class-imbalanced data.†

Estimator	*Symbol*	*Bias* $(\times 10^{-2})$	*Variance* $(\times 10^{-5})$	*MSE* $(\times 10^{-5})$
Baseline	$\hat{\alpha}_a$	0.000	1.960	1.960
Classify-and-count	$\hat{\alpha}^*$	−2.820	0.000	79,524
Subtracted-bias	$\hat{\alpha}_b$	−0.254	3.377	4.022
Misclassification	$\hat{\alpha}_p$	−0.003	3.587	3,587
Calibration	$\hat{\alpha}_c$	0.000	1.289	1.289

† Simulation results for $\alpha = 0.98$, $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$.

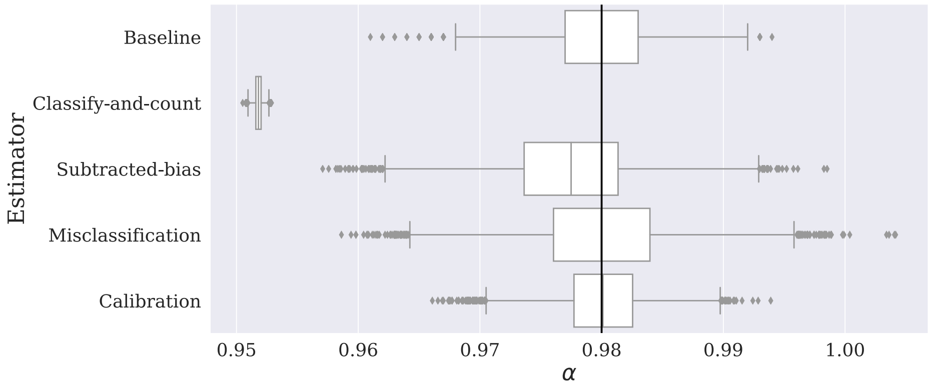


Fig. 2.2: The box plots show the sampling distribution of the estimators for α, where $\alpha = 0.98$, $p_{00} = 0.94$, $p_{11} = 0.97$, $n = 1000$ and $N = 3 \times 10^5$. The true value of α is highlighted by a vertical line.

points $(1 - \alpha, \alpha)$ and $(1, 1)$. As noted before, it implies that the classify-and-count estimator $\hat{\alpha}^*$ has small bias. Consequently, the subtracted-bias estimator $\hat{\alpha}_b$ has small bias as well. Thus, these two estimators will have the smallest MSE in the described region, whose size decreases as n increases. Fig. 2.3 visualises 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 data set increases. The biased methods, such as the classify-and-count estimator and the subtracted-bias estimator, perform well when the classification probabilities are large for the largest group.

As we have seen in both Table 2.2 and Table 2.3, the calibration estimator

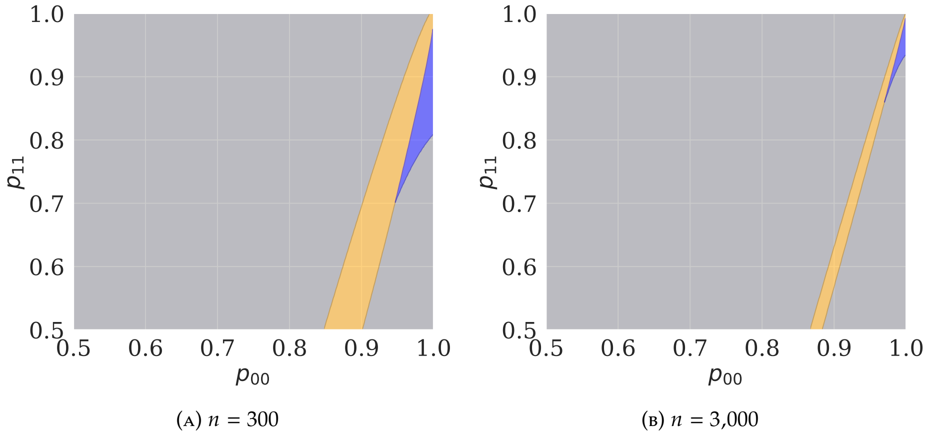
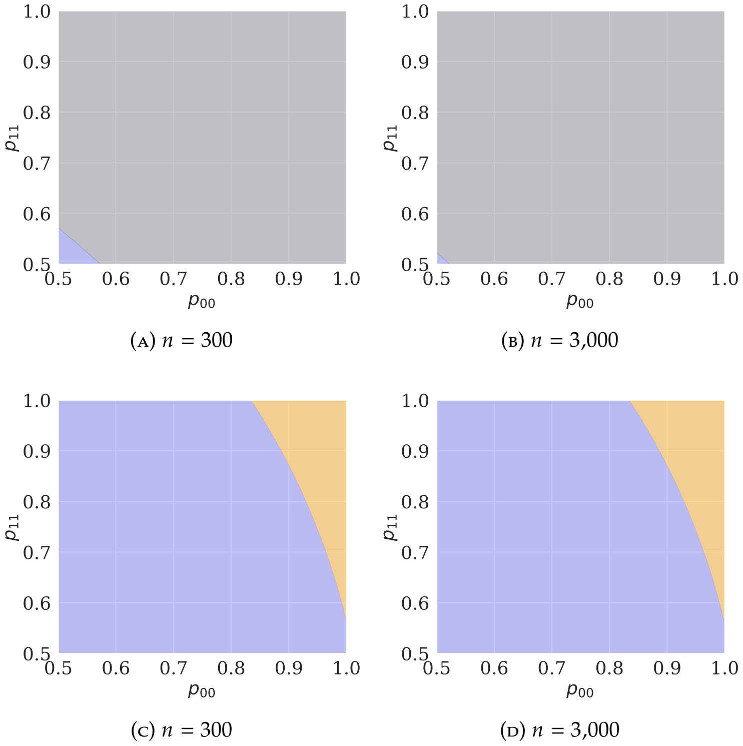


FIG. 2.3: For each coordinate (p_{00}, p_{11}), the depicted colour indicates which estimator has the smallest MSE, considering only the classify-and-count estimator (orange), the subtracted-bias estimator (blue) and the calibration estimator (grey). In panel (A), we have set $\alpha = 0.2$ and $n = 300$, whereas $\alpha = 0.2$ and $n = 3{,}000$ in the panel (B). The blue and orange regions are smaller in panel (B) compared with panel (A), 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.

$\hat{\alpha}_c$ competes with the baseline estimator in having the smallest MSE. In general, the calibration estimator will have smaller MSE if the classification probabilities p_{00} and p_{11} are larger, 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 smaller MSE 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 smaller MSE than the baseline estimator. Panels (A) and (B) in Fig. 2.4 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.

Table 2.2 and Table 2.3 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 (2.13) and (2.14). Panels (B) and (D) in Fig. 2.4 show, for $\alpha = 0.2$ and two different



(A) $n = 300$

(B) $n = 3{,}000$

(C) $n = 300$

(D) $n = 3{,}000$

Fig. 2.4: For each coordinate (p_{00}, p_{11}) the colour indicates which estimate has the smallest MSE, considering only the baseline estimator (blue), the calibration estimator (grey) and the misclassification estimator (orange). Panels (A) and (C) consider $\alpha = 0.2$ and $n = 300$, while panels (B) and (D) consider $\alpha = 0.2$ and $n = 3000$.

values of n, the curve in the (p_{00}, p_{11})-plane beyond which the misclassification estimator has smaller MSE than the baseline estimator. Observe that an increase in the size n of the test data set 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 of large values of p_{00} and p_{11}. In Theorem 2.4 it is proven that, for all possible combinations of α and sufficiently large n, the MSE of the calibration estimator is consistently smaller than that of the misclassification estimator.

Theorem 2.4. *Let $\widetilde{MSE}[\hat{\alpha}_p]$ and $\widetilde{MSE}[\hat{\alpha}_c]$ denote the approximate MSEs, up to terms of order $1/n$, of the misclassification estimator and the calibration estimator, respectively. It holds that:*

$$\widetilde{MSE}[\hat{\alpha}_p] - \widetilde{MSE}[\hat{\alpha}_c] = \frac{\Big[(1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})\Big]^2}{(p_{00}+p_{11}-1)^2\beta(1-\beta)}, \tag{2.18}$$

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

Proof. See Appendix 2.A. □

Thus, neglecting terms of order $1/n^2$ and higher, the result implies that the calibration estimator has a smaller MSE 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 $p_{00} + p_{11} \longrightarrow 1$. It is the result of the misclassification estimator having a singularity at $p_{00} + p_{11} = 1$ (see expression (2.14)), 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 $p_{00}+p_{11} \longrightarrow 1$, see Meertens, Diks, Van den Herik and Takes (2019) for a discussion and a possible solution.

2.4 Chapter conclusions

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 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 of α has smallest MSE. First, when dealing with small test data sets and rather poor algorithms, that is p_{00} and p_{11} both close to 0.5, the baseline estimator $\hat{\alpha}_a$ has the smallest 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 smallest MSE. As the size of the test data set increases, the size of that neighbourhood decreases. Third, in any other situation, the calibration estimator will have the smallest MSE. In practice, the test data set 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 generalised 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 (Moreno-Torres et al., 2012), is then inevitable. Consequently, we may no longer assume that the conditional distribution of the class label given the features in the test data set 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 smaller 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.

APPENDIX

2.A Theoretical derivations under assumption A1

This appendix contains the proofs of the theorems presented in Chapter 2. 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 which 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 as described in Section 2.2. 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 $\hat{p}_{00}$ defined in Section 2.2 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 $\mathbb{E}[f(X)]$ and $\mathbb{E}[g(X,Y)]$ for sufficiently differentiable functions f and g, we will insert the Taylor series for f and g at $x_0 = \mathbb{E}[X]$ and $y_0 = \mathbb{E}[Y]$ up to terms of order 2 and utilise the linearity of the expectation. Second, we will use the following conditional variance decomposition for the variance of a random variable X:

$$V(X) = \mathbb{E}[V(X \mid Y)] + V(\mathbb{E}[X \mid Y]). \tag{2.19}$$

The conditional variance decomposition follows from the tower property of conditional expectations Knottnerus (2003). Before we prove the theorems presented in the paper, we begin by proving the following lemma.

Lemma 2.1. *The variance of the estimator $\hat{p}_{11}$ for p_{11} estimated on the test set is given by*

$$V(\hat{p}_{11}) = \frac{p_{11}(1-p_{11})}{n\alpha}\left[1 + \frac{1-\alpha}{n\alpha}\right] + O\left(\frac{1}{n^3}\right). \tag{2.20}$$

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

$$V(\hat{p}_{00}) = \frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + O\left(\frac{1}{n^3}\right). \tag{2.21}$$

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

Proof of Lemma 2.1. We approximate the variance of $\hat{p}_{00}$ using the conditional variance decomposition and a second-order Taylor series. It follows that

$$\begin{aligned}
V(\hat{p}_{00}) &= V\left(\frac{n_{00}}{n_{0+}}\right) \\
&= E_{n_{0+}}\left[V\left(\frac{n_{00}}{n_{0+}} \mid n_{0+}\right)\right] + V_{n_{0+}}\left[\mathbb{E}\left(\frac{n_{00}}{n_{0+}} \mid n_{0+}\right)\right] \\
&= E_{n_{0+}}\left[\frac{1}{n_{0+}^2} V(n_{00} \mid n_{0+})\right] + V_{n_{0+}}\left[\frac{1}{n_{0+}}\mathbb{E}(n_{00} \mid n_{0+})\right] \\
&= E_{n_{0+}}\left[\frac{n_{0+}p_{00}(1-p_{00})}{n_{0+}^2}\right] + V_{n_{0+}}\left[\frac{n_{0+}p_{00}}{n_{0+}}\right] \\
&= E_{n_{0+}}\left[\frac{1}{n_{0+}}\right] p_{00}(1-p_{00}) + V_{n_{0+}}\left[p_{00}\right].
\end{aligned} \tag{2.22}$$

The second term of expression (2.22) is equal to 0. Hence, we find that

$$\begin{aligned} E_{n_{0+}}\left[\frac{1}{n_{0+}}\right] p_{00}(1-p_{00}) &= \left[\frac{1}{\mathbb{E}[n_{0+}]} + \frac{1}{2}\frac{2}{\mathbb{E}[n_{0+}]^3} \times V[n_{0+}]\right] p_{00}(1-p_{00}) + O\left(\frac{1}{n^3}\right) \\ &= \frac{p_{00}(1-p_{00})}{\mathbb{E}[n_{0+}]}\left[1 + \frac{V[n_{0+}]}{\mathbb{E}[n_{0+}]^2}\right] + O\left(\frac{1}{n^3}\right) \\ &= \frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + O\left(\frac{1}{n^3}\right). \end{aligned} \tag{2.23}$$

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

Finally, we use the analogue of (2.19) for covariances and we obtain that

$$\begin{aligned} C(\hat{p}_{11}, \hat{p}_{00}) &= C\left(\frac{n_{11}}{n_{1+}}, \frac{n_{00}}{n_{0+}}\right) \\ &= E_{n_{1+},n_{0+}}\left[C\left(\frac{n_{11}}{n_{1+}}, \frac{n_{00}}{n_{0+}} \mid n_{1+}, n_{0+}\right)\right] \\ &\quad + C_{n_{1+},n_{0+}}\left[\mathbb{E}\left(\frac{n_{11}}{n_{1+}} \mid n_{1+}, n_{0+}\right), \mathbb{E}\left(\frac{n_{00}}{n_{0+}} \mid n_{1+}, n_{0+}\right)\right] \\ &= E_{n_{1+},n_{0+}}\left[\frac{1}{n_{1+}n_{0+}} C(n_{11}, n_{00} \mid n_{1+}, n_{0+})\right] \\ &\quad + C_{n_{1+},n_{0+}}\left[\frac{1}{n_{1+}}\mathbb{E}(n_{11} \mid n_{1+}), \frac{1}{n_{0+}}\mathbb{E}(n_{00} \mid n_{0+})\right]. \end{aligned} \tag{2.24}$$

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 $C(n_{11}, n_{00} \mid n_{1+}, n_{0+}) = 0$. □

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

Subtracted-bias estimator

We will now prove the bias and variance approximations for the subtracted-bias estimator $\hat{\alpha}_b$ that was defined by Equation (2.9).

Proof of Theorem 2.1. The bias of $\hat{\alpha}_b$ is given by

$$\begin{aligned}
B(\hat{\alpha}_b) &= \mathbb{E}\left[\hat{\alpha}^\star - \hat{B}[\hat{\alpha}^\star]\right] - \alpha \\
&= \mathbb{E}[\hat{\alpha}^\star - \alpha] - \mathbb{E}\left[\hat{B}[\hat{\alpha}^\star]\right] \\
&= B[\hat{\alpha}^\star] - \mathbb{E}\left[\hat{B}[\hat{\alpha}^\star]\right] \\
&= \left[\alpha(p_{00} + p_{11} - 2) + (1 - p_{00})\right] - \mathbb{E}\left[\hat{\alpha}^\star(\hat{p}_{00} + \hat{p}_{11} - 2) + (1 - \hat{p}_{00})\right]. \qquad (2.25)
\end{aligned}$$

Because $\hat{\alpha}^*$ and $(\hat{p}_{00} + \hat{p}_{11} - 2)$ are assumed to be independent, the expectation of their product equals the product of their expectations. Subsequently, we obtain that

$$\begin{aligned}
B(\hat{\alpha}_b) &= \alpha(p_{00} + p_{11} - 2) + (1 - p_{00}) - \mathbb{E}[\hat{\alpha}^\star](p_{00} + p_{11} - 2) - (1 - p_{00}) \\
&= (\alpha - \mathbb{E}[\hat{\alpha}^\star])(p_{00} + p_{11} - 2) \\
&= B[\hat{\alpha}^\star](2 - p_{00} - p_{11}) \\
&= (1 - p_{00})(2 - p_{00} - p_{11}) - \alpha(p_{00} + p_{11} - 2)^2. \qquad (2.26)
\end{aligned}$$

This proves the formula for the bias of $\hat{\alpha}_b$ as estimator of α. To approximate the variance of $\hat{\alpha}_b$, we apply the conditional variance decomposition (2.19) conditional on $\hat{\alpha}^*$ and look at the two resulting terms separately. First, consider the expectation of the conditional variance

$$\begin{aligned}
\mathbb{E}\left[V(\hat{\alpha}_b \mid \hat{\alpha}^*)\right] &= \mathbb{E}\left[V(\hat{\alpha}^*(3 - \hat{p}_{00} - \hat{p}_{11}) - (1 - \hat{p}_{00}) \mid \hat{\alpha}^*)\right] \\
&= \mathbb{E}\Big[V(\hat{\alpha}^*(3 - \hat{p}_{00} - \hat{p}_{11}) \mid \hat{\alpha}^*) + V(1 - \hat{p}_{00} \mid \hat{\alpha}^*) \\
&\qquad - 2C(\hat{\alpha}^*(3 - \hat{p}_{00} - \hat{p}_{11}), 1 - \hat{p}_{00} \mid \hat{\alpha}^*)\Big] \\
&= \mathbb{E}\Big[(\hat{\alpha}^*)^2 V(3 - \hat{p}_{00} - \hat{p}_{11} \mid \hat{\alpha}^*) + V(1 - \hat{p}_{00} \mid \hat{\alpha}^*) \\
&\qquad - 2\hat{\alpha}^* C(3 - \hat{p}_{00} - \hat{p}_{11}, 1 - \hat{p}_{00} \mid \hat{\alpha}^*)\Big] \\
&= \mathbb{E}\left[(\hat{\alpha}^*)^2\left[V(\hat{p}_{00}) + V(\hat{p}_{11})\right] + V(\hat{p}_{00}) - 2\hat{\alpha}^* V(\hat{p}_{00})\right] \\
&= \mathbb{E}\left[(\hat{\alpha}^*)^2\right]\left[V(\hat{p}_{00}) + V(\hat{p}_{11})\right] + V(\hat{p}_{00}) - 2\,\mathbb{E}\left[\hat{\alpha}^*\right] V(\hat{p}_{00}). \qquad (2.27)
\end{aligned}$$

In the penultimate line, we used that $C(\hat{p}_{11}, \hat{p}_{00}) = 0$. The second moment $\mathbb{E}\left[(\hat{\alpha}^*)^2\right]$ can be written as $\mathbb{E}\left[\hat{\alpha}^*\right]^2 + V(\hat{\alpha}^*)$. Because $V(\hat{\alpha}^*)$ is of order $1/N$, it can be neglected compared to $\mathbb{E}\left[\hat{\alpha}^*\right]^2$, which is of order 1. In particular, the expectation of the conditional variance can be expressed as

$$\begin{aligned}\mathbb{E}\left[V(\hat{\alpha}_b \mid \hat{\alpha}^*)\right] &= \mathbb{E}\left[(\hat{\alpha}^*)\right]^2 \left[V(\hat{p}_{00}) + V(\hat{p}_{11})\right] + V(\hat{p}_{00}) - 2\,\mathbb{E}\left[\hat{\alpha}^*\right] V(\hat{p}_{00}) + O\left(\frac{1}{N}\right) \\ &= V(\hat{p}_{00}) \left[\mathbb{E}\left[\hat{\alpha}^*\right] - 1\right]^2 + V(\hat{p}_{11})\,\mathbb{E}\left[\hat{\alpha}^*\right]^2 + O\left(\frac{1}{N}\right). \qquad (2.28)\end{aligned}$$

Next, the variance of the conditional expectation can be rewritten as

$$\begin{aligned}V\left[\mathbb{E}(\hat{\alpha}_b \mid \hat{\alpha}^*)\right] &= V\left[\mathbb{E}(\hat{\alpha}^*(3 - \hat{p}_{00} - \hat{p}_{11}) - (1 - \hat{p}_{00}) \mid \hat{\alpha}^*)\right] \\ &= V\left[\hat{\alpha}^*\,\mathbb{E}(3 - \hat{p}_{00} - \hat{p}_{11} \mid \hat{\alpha}^*) - \mathbb{E}(1 - \hat{p}_{00} \mid \hat{\alpha}^*)\right] \\ &= V(\hat{\alpha}^*)(3 - p_{00} - p_{11})^2. \qquad (2.29)\end{aligned}$$

Because $V(\hat{\alpha}^*)$ is of order $1/N$, it can be neglected in the final formula. Furthermore, the variances of $\hat{p}_{00}$ and $\hat{p}_{11}$ can be computed using the result from Lemma 2.1, giving

$$\begin{aligned}V(\hat{\alpha}_b) = {} & \frac{\left[\alpha(p_{00} + p_{11} - 1) - p_{00}\right]^2 p_{00}(1 - p_{00})}{n(1 - \alpha)} \left[1 + \frac{\alpha}{n(1 - \alpha)}\right] \\ & + \frac{\left[\alpha(p_{00} + p_{11} - 1) + (1 - p_{00})\right]^2 p_{11}(1 - p_{11})}{n\alpha} \left[1 + \frac{1 - \alpha}{n\alpha}\right] \\ & + O\left(\max\left[\frac{1}{n^3}, \frac{1}{N}\right]\right). \qquad (2.30)\end{aligned}$$

This concludes the proof of Theorem 2.1. □

Misclassification estimator

We will now prove the bias and variance approximations for the misclassification estimator $\hat{\alpha}_p$ as defined by Equation (2.12).

Proof of Theorem 2.2. Under the assumption that $\hat{\alpha}^*$ is distributed independently of $(\hat{p}_{00}, \hat{p}_{11})$, it holds that

$$\begin{aligned}\mathbb{E}(\hat{\alpha}_p) &= \mathbb{E}\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) + \mathbb{E}\left[\mathbb{E}\left(\frac{\hat{\alpha}^*}{\hat{p}_{00}+\hat{p}_{11}-1}\,\middle|\,\hat{\alpha}^*\right)\right] \\ &= \mathbb{E}\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) + \mathbb{E}(\hat{\alpha}^*)\,\mathbb{E}\left(\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right). \qquad (2.31)\end{aligned}$$

$\mathbb{E}(\hat{\alpha}^*)$ is known from (2.4). To evaluate the other two expectations, we use a second-order Taylor series approximation. The first- and second-order partial derivatives of $f(x, y) = 1/(x + y - 1)$ are given by

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = \frac{-1}{(x+y-1)^2}, \qquad (2.32)$$

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial y^2} = \frac{2}{(x+y-1)^3}. \qquad (2.33)$$

The partial derivatives of $g(x, y) = (x - 1)/(x + y - 1) = 1 - [y/(x + y - 1)]$ are equal to the expressions

$$\frac{\partial g}{\partial x} = \frac{y}{(x+y-1)^2}, \qquad \frac{\partial g}{\partial y} = \frac{-(x-1)}{(x+y-1)^2}, \qquad (2.34)$$

$$\frac{\partial^2 g}{\partial x^2} = \frac{-2y}{(x+y-1)^3}, \qquad \frac{\partial^2 g}{\partial y^2} = \frac{2(x-1)}{(x+y-1)^3}. \qquad (2.35)$$

Now also using that $C(\hat{p}_{11}, \hat{p}_{00}) = 0$, we obtain for the first expectation that

$$\begin{aligned}\mathbb{E}\left(\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) &= \frac{1}{p_{00}+p_{11}-1} + \frac{V(\hat{p}_{00}) + V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^3} + O(n^{-2}) \\ &= \frac{1}{p_{00}+p_{11}-1}\left[1 + \frac{\frac{p_{00}(1-p_{00})}{n(1-\alpha)} + \frac{p_{11}(1-p_{11})}{n\alpha}}{(p_{00}+p_{11}-1)^2}\right] + O(n^{-2}). \qquad (2.36)\end{aligned}$$

Here, we have included only the first term of the approximations to $V(\hat{p}_{00})$ and $V(\hat{p}_{11})$ from Lemma 2.1, since this suffices to approximate the bias up to terms of order $O(1/n)$. Similarly, for the second expectation we obtain that

$$\begin{aligned}\mathbb{E}\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right) &= \frac{p_{00}-1}{p_{00}+p_{11}-1} + \frac{(p_{00}-1)V(\hat{p}_{11}) - p_{11}V(\hat{p}_{00})}{(p_{00}+p_{11}-1)^3} + O(n^{-2}) \\ &= \frac{p_{00}-1}{p_{00}+p_{11}-1}\left[1 + p_{11}\frac{\frac{1-p_{11}}{n\alpha} + \frac{p_{00}}{n(1-\alpha)}}{(p_{00}+p_{11}-1)^2}\right] + O(n^{-2}). \qquad (2.37)\end{aligned}$$

Using (2.31), (2.4), (2.36), and (2.37), we conclude that

$$\begin{aligned}\mathbb{E}(\hat{\alpha}_p) = {}& \frac{\alpha(p_{00}+p_{11}-1)-(p_{00}-1)}{p_{00}+p_{11}-1}\left[1 + \frac{\frac{p_{00}(1-p_{00})}{n(1-\alpha)} + \frac{p_{11}(1-p_{11})}{n\alpha}}{(p_{00}+p_{11}-1)^2}\right] \\ &+ \frac{p_{00}-1}{p_{00}+p_{11}-1}\left[1 + p_{11}\frac{\frac{1-p_{11}}{n\alpha} + \frac{p_{00}}{n(1-\alpha)}}{(p_{00}+p_{11}-1)^2}\right] + O\left(\frac{1}{n^2}\right). \qquad (2.38)\end{aligned}$$

From this, it follows that an approximation to the bias of $\hat{\alpha}_p$ that is correct up to terms of order $O(1/n)$ is given by

$$\begin{aligned}B(\hat{\alpha}_p) = {}& \frac{\alpha(p_{00}+p_{11}-1)-(p_{00}-1)}{n(p_{00}+p_{11}-1)^3}\left[\frac{p_{00}(1-p_{00})}{1-\alpha} + \frac{p_{11}(1-p_{11})}{\alpha}\right] \\ &+ \frac{(p_{00}-1)p_{11}}{n(p_{00}+p_{11}-1)^3}\left[\frac{1-p_{11}}{\alpha} + \frac{p_{00}}{1-\alpha}\right] + O\left(\frac{1}{n^2}\right). \qquad (2.39)\end{aligned}$$

By expanding the products in this expression and combining similar terms, the expression can be simplified to

$$B(\hat{\alpha}_p) = \frac{p_{11}(1-p_{11}) - p_{00}(1-p_{00})}{n(p_{00}+p_{11}-1)^2} + O\left(\frac{1}{n^2}\right). \qquad (2.40)$$

Finally, using the identity $p_{11}(1-p_{11}) - p_{00}(1-p_{00}) = (p_{00}+p_{11}-1)(p_{00}-p_{11})$, we obtain the required result for $B(\hat{\alpha}_p)$.

To approximate the variance of $\hat{\alpha}_p$, we apply the conditional variance decomposition conditional on $\hat{\alpha}^*$ and look at the two resulting terms separately. First, consider the variance of the conditional expectation

$$\begin{aligned} V\left[\mathbb{E}(\hat{\alpha}_p \mid \hat{\alpha}^*)\right] &= V\left[\mathbb{E}\left(\hat{\alpha}^* \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} + \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^*\right)\right] \\ &= V\left[\hat{\alpha}^* \frac{1}{p_{00}+p_{11}-1}\right] \\ &= \frac{1}{(p_{00}+p_{11}-1)^2} V\left[\hat{\alpha}^*\right] = O\left(\frac{1}{N}\right), \end{aligned} \tag{2.41}$$

where in the last line we used expression (2.6). The factor $1/(p_{00}+p_{11}-1)^2$ can become arbitrarily large in the limit $p_{00}+p_{11} \to 1$. We will show below that this same factor also occurs in the lower-order terms of $V(\hat{\alpha}_p)$. Hence, the relative contribution of (2.41) remains negligible even in the limit $p_{00}+p_{11} \to 1$.

Next, we compute the expectation of the conditional variance

$$\begin{aligned} \mathbb{E}\left[V(\hat{\alpha}_p \mid \hat{\alpha}^*)\right] &= \mathbb{E}\left[V\left(\hat{\alpha}^* \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} + \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^\star\right)\right] \\ &= \mathbb{E}\left[V\left(\hat{\alpha}^* \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \alpha^\star\right) + V\left(\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^\star\right)\right. \\ &\quad \left. + 2C\left(\hat{\alpha}^* \frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1} \mid \hat{\alpha}^\star\right)\right] \\ &= \mathbb{E}\left[(\hat{\alpha}^*)^2\right] V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] + V\left[\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\ &\quad + 2\,\mathbb{E}\left[\hat{\alpha}^\star\right] C\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\ &= \mathbb{E}\left[\hat{\alpha}^\star\right]^2 \left[1 + O\left(\frac{1}{N}\right)\right] V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] + V\left[\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] \\ &\quad + 2\,\mathbb{E}\left[\hat{\alpha}^\star\right] C\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right]. \end{aligned} \tag{2.42}$$

To approximate the variance and covariance terms, we use a first-order Taylor series. Using the partial derivatives in (2.32) and (2.34), we obtain that

$$V\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = \frac{V(\hat{p}_{00})+V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^4} + O(n^{-2}) \tag{2.43}$$

$$V\left[\frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = \frac{V(\hat{p}_{00})(p_{11})^2}{(p_{00}+p_{11}-1)^4} + \frac{V(\hat{p}_{11})(1-p_{00})^2}{(p_{00}+p_{11}-1)^4} + O(n^{-2}) \tag{2.44}$$

$$C\left[\frac{1}{\hat{p}_{00}+\hat{p}_{11}-1}, \frac{\hat{p}_{00}-1}{\hat{p}_{00}+\hat{p}_{11}-1}\right] = \frac{V(\hat{p}_{00})(-p_{11})}{(p_{00}+p_{11}-1)^4} + \frac{V(\hat{p}_{11})(p_{00}-1)}{(p_{00}+p_{11}-1)^4} + O(n^{-2}). \tag{2.45}$$

Substituting these terms into expression (2.42) and accounting for expression (2.41) yields the following

$$\begin{aligned} V(\hat{\alpha}_p) = & \frac{V(\hat{p}_{00})\left[\mathbb{E}\left[\hat{\alpha}^{\star}\right]^2 - 2p_{11}\,\mathbb{E}\left[\hat{\alpha}^{\star}\right] + p_{11}^2\right]}{(p_{00}+p_{11}-1)^4} \\ & + \frac{V(\hat{p}_{11})\left[\mathbb{E}\left[\hat{\alpha}^{\star}\right]^2 - 2(1-p_{00})\,\mathbb{E}\left[\hat{\alpha}^{\star}\right] + (1-p_{00})^2\right]}{(p_{00}+p_{11}-1)^4} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right) \\ = & \frac{V(\hat{p}_{00})\left[\mathbb{E}\left[\hat{\alpha}^{\star}\right] - p_{11}\right]^2}{(p_{00}+p_{11}-1)^4} + \frac{V(\hat{p}_{11})\left[\mathbb{E}\left[\hat{\alpha}^{\star}\right] - (1-p_{00})\right]^2}{(p_{00}+p_{11}-1)^4} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right) \\ = & \frac{V(\hat{p}_{00})(1-\alpha)^2}{(p_{00}+p_{11}-1)^2} + \frac{V(\hat{p}_{11})\alpha^2}{(p_{00}+p_{11}-1)^2} + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right). \end{aligned} \tag{2.46}$$

Finally, inserting the expressions for $V(\hat{p}_{00})$ and $V(\hat{p}_{11})$ from Lemma 2.1 yields

$$\begin{aligned} V(\hat{\alpha}_p) = & \frac{\frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1+\frac{\alpha}{n(1-\alpha)}\right](1-\alpha)^2}{(p_{00}+p_{11}-1)^2} + \frac{\frac{p_{11}(1-p_{11})}{n\alpha}\left[1+\frac{1-\alpha}{n\alpha}\right]\alpha^2}{(p_{00}+p_{11}-1)^2} \\ & + O\left(\max\left[\frac{1}{n^2}, \frac{1}{N}\right]\right), \end{aligned} \tag{2.47}$$

from which expression (2.14) follows. This concludes the proof of Theorem 2.2. □

Calibration estimator

We will now prove the bias and variance approximations for the calibration estimator $\hat{\alpha}_c$ that was defined by Equation (2.15).

Proof of Theorem 2.3. To compute the expected value of $\hat{\alpha}_c$, we first compute its expectation conditional on the 4-vector $\boldsymbol{N} = (N_{00}, N_{01}, N_{10}, N_{11})$ and find that

$$\begin{aligned}
\mathbb{E}(\hat{\alpha}_c \mid \boldsymbol{N}) &= \mathbb{E}\left[\hat{\alpha}^* \frac{n_{11}}{n_{+1}} + (1-\hat{\alpha}^*)\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\
&= \hat{\alpha}^*\,\mathbb{E}\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] + (1-\hat{\alpha}^*)\,\mathbb{E}\left[\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\
&= \hat{\alpha}^*\,\mathbb{E}\left[\mathbb{E}\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right] \\
&\qquad + (1-\hat{\alpha}^*)\,\mathbb{E}\left[\mathbb{E}\left(\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}, n_{+0}\right) \mid \boldsymbol{N}\right] \\
&= \frac{N_{+1}}{N}\,\mathbb{E}\left[\frac{1}{n_{+1}} n_{+1} \frac{N_{11}}{N_{+1}} \mid \boldsymbol{N}\right] + \frac{N_{+0}}{N}\,\mathbb{E}\left[\frac{1}{n_{+0}} n_{+0} \frac{N_{10}}{N_{+0}} \mid \boldsymbol{N}\right] \\
&= \frac{N_{11}}{N} + \frac{N_{10}}{N} \\
&= \frac{N_{1+}}{N} = \alpha. \qquad (2.48)
\end{aligned}$$

The tower property of conditional expectations implies $\mathbb{E}[\hat{\alpha}_c] = \mathbb{E}\left[\mathbb{E}(\hat{\alpha}_c \mid \boldsymbol{N})\right] = \alpha$. This proves that $\hat{\alpha}_c$ is an unbiased estimator of α.

To compute the variance of $\hat{\alpha}_c$, we use the conditional variance decomposition, again conditioning on the 4-vector $\boldsymbol{N}$. We remark that N_{0+} and N_{1+} are deterministic values, but that N_{+0} and N_{+1} are random variables. As shown above in Equation (2.48), the conditional expectation is deterministic, hence it has no variance, i.e., $V(\mathbb{E}[\hat{\alpha}_c \mid \boldsymbol{N}]) = 0$. The conditional variance decomposition then simplifies to

$$V(\hat{\alpha}_c) = \mathbb{E}\left[V(\hat{\alpha}_c \mid \boldsymbol{N})\right]. \qquad (2.49)$$

The conditional variance $V(\hat{\alpha}_c \mid \boldsymbol{N})$ can be written as

$$\begin{aligned} V[\hat{\alpha}_c \mid \boldsymbol{N}] &= V\left[\hat{\alpha}^* \frac{n_{11}}{n_{+1}} + (1-\hat{\alpha}^*)\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\ &= (\hat{\alpha}^*)^2 V\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] + (1-\hat{\alpha}^*)^2 V\left[\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] \\ &\quad + 2\hat{\alpha}^*(1-\hat{\alpha}^*) C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right]. \end{aligned} \tag{2.50}$$

We will consider these terms separately. First, the variance of n_{11}/n_{+1} can be computed by applying an additional conditional variance decomposition as

$$V\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] = V\left[\mathbb{E}\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right] + \mathbb{E}\left[V\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right]. \tag{2.51}$$

The first term is zero, which follows from

$$\begin{aligned} V\left[\mathbb{E}\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right)\right] &= V\left[\frac{1}{n_{+1}} \mathbb{E}(n_{11} \mid \boldsymbol{N}, n_{+1}) \mid \boldsymbol{N}\right] \\ &= V\left[\frac{1}{n_{+1}} n_{+1} \frac{N_{11}}{N_{+1}} \mid \boldsymbol{N}\right] \\ &= V\left[\frac{N_{11}}{N_{+1}} \mid \boldsymbol{N}\right] = 0. \end{aligned} \tag{2.52}$$

For the second term, under the assumption that $n \ll N$, we find that

$$\begin{aligned} \mathbb{E}\left[V\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+1}\right) \mid \boldsymbol{N}\right] &= \mathbb{E}\left[\frac{1}{n_{+1}^2} V(n_{11} \mid \boldsymbol{N}, n_{+1}) \mid \boldsymbol{N}\right] \\ &= \mathbb{E}\left[\frac{1}{n_{+1}^2} n_{+1} \frac{N_{11}}{N_{+1}} (1 - \frac{N_{11}}{N_{+1}}) \mid \boldsymbol{N}\right] \\ &= \mathbb{E}\left[\frac{1}{n_{+1}} \mid \boldsymbol{N}\right] \frac{N_{11} N_{01}}{N_{+1}^2}. \end{aligned} \tag{2.53}$$

The expectation of $\frac{1}{n_{+1}}$ can be approximated with a second-order Taylor series

$$\begin{aligned} V\left[\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}\right] &= \left[\frac{1}{\mathbb{E}[n_{+1} \mid \boldsymbol{N}]} + \frac{1}{2}\frac{2}{\mathbb{E}[n_{+1} \mid \boldsymbol{N}]^3} V\left[n_{+1} \mid \boldsymbol{N}\right]\right] \frac{N_{11}N_{01}}{N_{+1}^2} + O(n^{-3}) \\ &= \frac{1}{\mathbb{E}[n_{+1} \mid \boldsymbol{N}]}\left[1 + \frac{V\left[n_{+1} \mid \boldsymbol{N}\right]}{\mathbb{E}[n_{+1} \mid \boldsymbol{N}]^2}\right] \frac{N_{11}N_{01}}{N_{+1}^2} + O(n^{-3}) \\ &= \frac{1}{n\hat{\alpha}^*}\left[1 + \frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right] \frac{N_{11}N_{01}}{N_{+1}^2} + O(n^{-3}). \end{aligned} \tag{2.54}$$

The variance of n_{10}/n_{+0} can be approximated in the same way, which yields

$$V\left[\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] = \frac{1}{n(1-\hat{\alpha}^*)}\left[1 + \frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right] \frac{N_{00}N_{10}}{N_{+0}^2} + O(n^{-3}). \tag{2.55}$$

Finally, it can be shown that the covariance in the final term is equal to zero by

$$\begin{aligned} C\left[\frac{n_{11}}{n_{+1}}, \frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}\right] &= \mathbb{E}\left[C\left(\frac{n_{11}}{n_{+1}}, \frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &\quad + C\left[\mathbb{E}\left(\frac{n_{11}}{n_{+1}} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right), \mathbb{E}\left(\frac{n_{10}}{n_{+0}} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &= \mathbb{E}\left[\frac{1}{n_{+0}n_{+1}} C\left(n_{11}, n_{10} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &\quad + C\left[\frac{1}{n_{+1}}\mathbb{E}\left(n_{11} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right), \frac{1}{n_{+0}}\mathbb{E}\left(n_{10} \mid \boldsymbol{N}, n_{+0}, n_{+1}\right) \mid \boldsymbol{N}\right] \\ &= 0 + C\left[\frac{1}{n_{+1}} n_{+1} \frac{N_{11}}{N_{+1}}, \frac{1}{n_{+0}} n_{+0} \frac{N_{10}}{N_{+0}} \mid \boldsymbol{N}\right] = 0. \end{aligned} \tag{2.56}$$

Combining expressions (2.54), (2.55) and (2.56) with (2.50) gives

$$\begin{aligned} V[\hat{\alpha}_c \mid \boldsymbol{N}] &= \frac{N_{+1}^2}{N^2}\frac{1}{n\hat{\alpha}^*}\left[1+\frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right]\frac{N_{11}N_{01}}{N_{+1}^2} \\ &\quad + \frac{N_{+0}^2}{N^2}\frac{1}{n(1-\hat{\alpha}^*)}\left[1+\frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right]\frac{N_{00}N_{10}}{N_{+0}^2} + O(n^{-3}) \\ &= \frac{1}{n\hat{\alpha}^*}\left[1+\frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right]\frac{N_{11}N_{01}}{N^2} \\ &\quad + \frac{1}{n(1-\hat{\alpha}^*)}\left[1+\frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right]\frac{N_{00}N_{10}}{N^2} + O(n^{-3}). \end{aligned} \tag{2.57}$$

Recall from Formula (2.49) that $V[\hat{\alpha}_c] = \mathbb{E}[V[\hat{\alpha}_c \mid \boldsymbol{N}]] = \mathbb{E}[\mathbb{E}[V[\hat{\alpha}_c \mid \boldsymbol{N}] \mid N_{+1}]]$. Hence,

$$\begin{aligned} V[\hat{\alpha}_c] = \mathbb{E}\Bigg[&\frac{1}{n\hat{\alpha}^*}\left(1+\frac{1-\hat{\alpha}^*}{n\hat{\alpha}^*}\right)\mathbb{E}\left(\frac{N_{11}N_{01}}{N^2} \mid N_{+1}\right) \\ &+\frac{1}{n(1-\hat{\alpha}^*)}\left(1+\frac{\hat{\alpha}^*}{n(1-\hat{\alpha}^*)}\right)\mathbb{E}\left(\frac{N_{00}N_{10}}{N^2} \mid N_{+1}\right)\Bigg] + O(n^{-3}). \end{aligned} \tag{2.58}$$

To evaluate the expectations in this expression, we observe that, conditional on the column total N_{+1}, N_{11} is distributed as $Bin(N_{+1}, c_{11})$, where c_{11} is a calibration probability as defined in Section 2.2.5. Hence,

$$\mathbb{E}[N_{11} \mid N_{+1}] = N_{+1}c_{11} = \frac{N_{+1}\alpha p_{11}}{(1-\alpha)(1-p_{00})+\alpha p_{11}} \tag{2.59}$$

$$V[N_{11} \mid N_{+1}] = N_{+1}c_{11}(1-c_{11}). \tag{2.60}$$

Similarly, since $N = N_{+1} + N_{+0}$ is fixed,

$$\mathbb{E}[N_{00} \mid N_{+1}] = N_{+0}c_{00} = \frac{N_{+0}(1-\alpha)p_{00}}{(1-\alpha)p_{00}+\alpha(1-p_{11})} \tag{2.61}$$

$$V[N_{00} \mid N_{+1}] = N_{+0}c_{00}(1-c_{00}). \tag{2.62}$$

Using these results, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\frac{N_{11}N_{01}}{N^2} \mid N_{+1}\right] &= \frac{1}{N^2}\,\mathbb{E}\left[N_{11}N_{01} \mid N_{+1}\right] \\
&= \frac{1}{N^2}\,\mathbb{E}\left[N_{11}(N_{+1} - N_{11}) \mid N_{+1}\right] \\
&= \frac{1}{N^2}\left[N_{+1}\,\mathbb{E}\left[N_{11} \mid N_{+1}\right] - \mathbb{E}\left[N_{11}^2 \mid N_{+1}\right]\right] \\
&= \frac{1}{N^2}\left[N_{+1}\,\mathbb{E}\left[N_{11} \mid N_{+1}\right] - V\left[N_{11} \mid N_{+1}\right] - \mathbb{E}\left[N_{11} \mid N_{+1}\right]^2\right] \\
&= \frac{1}{N^2}\left[N_{+1}^2 c_{11} - N_{+1}c_{11}(1 - c_{11}) - N_{+1}^2 c_{11}^2\right] \\
&= \frac{N_{+1}^2}{N^2} c_{11}(1 - c_{11}) + O\left(\frac{1}{N}\right), \qquad (2.63)
\end{aligned}$$

and similarly

$$\mathbb{E}\left[\frac{N_{00}N_{10}}{N^2} \mid N_{+1}\right] = \frac{N_{+0}^2}{N^2} c_{00}(1 - c_{00}) + O\left(\frac{1}{N}\right). \qquad (2.64)$$

Substituting expressions (2.63) and (2.64) into expression (2.58) and noting that $N_{+1}^2/N^2 = (\hat{\alpha}^*)^2$ and $N_{+0}^2/N^2 = (1 - \hat{\alpha}^*)^2$, we find that

$$\begin{aligned}
V[\hat{\alpha}_c] = \mathbb{E}\Bigg[&\frac{\hat{\alpha}^*}{n}\left(1 + \frac{1 - \hat{\alpha}^*}{n\hat{\alpha}^*}\right) c_{11}(1 - c_{11}) \\
&+ \frac{1 - \hat{\alpha}^*}{n}\left(1 + \frac{\hat{\alpha}^*}{n(1 - \hat{\alpha}^*)}\right) c_{00}(1 - c_{00})\Bigg] + O\left(\max\left[\frac{1}{n^3}, \frac{1}{Nn}\right]\right) \\
= &\left[\frac{\mathbb{E}(\hat{\alpha}^*)}{n} + \frac{1 - \mathbb{E}(\hat{\alpha}^*)}{n^2}\right] c_{11}(1 - c_{11}) \\
&+ \left[\frac{1 - \mathbb{E}(\hat{\alpha}^*)}{n} + \frac{\mathbb{E}(\hat{\alpha}^*)}{n^2}\right] c_{00}(1 - c_{00}) + O\left(\max\left[\frac{1}{n^3}, \frac{1}{Nn}\right]\right). \qquad (2.65)
\end{aligned}$$

Finally, substituting the expressions for $\mathbb{E}(\hat{\alpha}^*)$ from (2.4) and the expressions for c_{11} and c_{00} from (2.59) and (2.61), the desired expression (2.17) is obtained. This concludes the proof of Theorem 2.3. □

Comparing mean squared errors

To conclude, we present the proof of Theorem 2.4, which essentially shows that the MSE (up to and including terms of order $1/n$) of the calibration estimator is smaller than that of the misclassification estimator.

Proof of Theorem 2.4. Recall that the bias of $\hat{\alpha}_p$ as an estimator of α is given by

$$B\left[\hat{\alpha}_p\right] = \frac{p_{00} - p_{11}}{n(p_{00} + p_{11} - 1)} + O\left(\frac{1}{n^2}\right). \tag{2.66}$$

Hence, $(B\left[\hat{\alpha}_p\right])^2 = O(1/n^2)$ is not relevant for $\widetilde{MSE}[\hat{\alpha}_p]$. It follows that $\widetilde{MSE}[\hat{\alpha}_p]$ is equal to the variance of $\hat{\alpha}_p$ up to order $1/n$. From (2.14) we obtain that

$$\widetilde{MSE}[\hat{\alpha}_p] = \frac{1}{n}\left[\frac{(1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})}{(p_{00}+p_{11}-1)^2}\right]. \tag{2.67}$$

Recall that $\hat{\alpha}_c$ is an unbiased estimator of α, i.e., $B[\hat{\alpha}_c] = 0$. Also recall the notation $\beta = (1-\alpha)(1-p_{00}) + \alpha p_{11}$. It follows from (2.17) that the variance, and hence the MSE, of $\hat{\alpha}_c$ up to terms of order $1/n$ can be written as

$$\begin{aligned}\widetilde{MSE}[\hat{\alpha}_c] &= \frac{1}{n}\left[\beta\frac{\alpha p_{11}}{\beta}\left(1 - \frac{\alpha p_{11}}{\beta}\right) + (1-\beta)\frac{(1-\alpha)p_{00}}{1-\beta}\left(1 - \frac{(1-\alpha)p_{00}}{1-\beta}\right)\right] \\ &= \frac{\alpha(1-\alpha)}{n}\left[\frac{(1-p_{00})p_{11}}{\beta} + \frac{p_{00}(1-p_{11})}{1-\beta}\right]. \end{aligned} \tag{2.68}$$

To prove Expression (2.18), first note that

$$\frac{(1-p_{00})p_{11}}{\beta} + \frac{p_{00}(1-p_{11})}{1-\beta} = \frac{(1-p_{00})p_{11} + \beta(p_{00}-p_{11})}{\beta(1-\beta)}. \tag{2.69}$$

The numerator of this equation can be rewritten as

$$\begin{aligned}&(1-p_{00})p_{11} + \beta(p_{00}-p_{11}) \\ &= (1-p_{00})p_{11} + (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{00}p_{11} - (1-\alpha)(1-p_{00})p_{11} - \alpha p_{11}^2 \\ &= (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{00}p_{11} + \alpha(1-p_{00})p_{11} - \alpha p_{11}^2 \\ &= (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11}). \end{aligned} \tag{2.70}$$

Note that the obtained expression is equal to the numerator of Expression (2.67). Write $T = (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$ for that expression. It follows that

$$\begin{aligned}
&\widetilde{MSE}[\hat{\alpha}_p] - \widetilde{MSE}[\hat{\alpha}_c] \\
&= \frac{T}{n(p_{00}+p_{11}-1)^2} - \frac{T\alpha(1-\alpha)}{n\beta(1-\beta)} \\
&= \frac{T}{n(p_{00}+p_{11}-1)^2\beta(1-\beta)}\Big[\beta(1-\beta) - \alpha(1-\alpha)(p_{00}+p_{11}-1)^2\Big]. \qquad (2.71)
\end{aligned}$$

Rewriting the second factor in the last expression gives

$$\begin{aligned}
&\beta(1-\beta) - \alpha(1-\alpha)(p_{00}+p_{11}-1)^2 \\
&= (1-\alpha)^2 p_{00}(1-p_{00}) + \alpha(1-\alpha)\Big((1-p_{00})(1-p_{11}) + p_{00}p_{11}\Big) + \alpha^2 p_{11}(1-p_{11}) \\
&\qquad - \alpha(1-\alpha)(p_{00}+p_{11}-1)^2 \\
&= (1-\alpha)^2 p_{00}(1-p_{00}) + \alpha(1-\alpha)\Big(p_{00}(1-p_{00}) + p_{11}(1-p_{11})\Big) + \alpha^2 p_{11}(1-p_{11}) \\
&= (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11}) \\
&= T. \qquad (2.72)
\end{aligned}$$

This concludes the proof of Theorem 2.4. □

CHAPTER 3

PRIOR PROBABILITY SHIFT

3.1 Introduction

In recent years, many national statistical institutes (NSIs) have experimented with supervised machine learning algorithms with the purpose of producing new or improved official statistics. Beck et al. (2018) provide a list of 136 machine learning projects at NSIs in 25 countries. In many projects, machine learning was used for classification (78) or for imputation (22). The results of these machine learning projects are promising and therefore currently seen as a paradigm shift in official statistics, in which model-based statistics are widely embraced (De Broe et al., 2020).

The quality of the statistical output is a key challenge when employing classification algorithms for producing official statistics. Output quality is a fundamental component in any quality framework for official statistics, see, e.g., the OECD quality framework (OECD, 2011) and the Regulation on European Statistics (European Commission, 2009) translated into the European Statistics Code of Practice (Eurostat, 2017). When using classification algorithms for official statistics, the output quality ought to be measured using the MSE of the statistical output (Buelens et al., 2016).

In the machine learning literature, the accuracy of classification algorithms is measured at the level of individual data points. Interestingly, the algorithmic accuracy at the level of *individual* data points differs fundamentally from the accuracy (at the *population* level) of the (aggregated) statistical output of classification algorithms (Forman, 2005). In fact, classification algorithms that have high algorithmic accuracy might still produce highly biased statistical output. This is referred to as *misclassification bias*. It is a type of bias that is commonly overlooked or neglected by statisticians of all time (González et al., 2017; Schwartz, 1985).

After many years of persistent research, a rich body of statistical literature on misclassification bias is readily available. Misclassification bias occurs in general when dealing with measurement errors in categorical data. The work by Bross (1954) is usually referred to as the first publication to discuss the problem of misclassification bias. Other significant contributions to the literature on misclassification bias include the work by Tenenbein (1970) and the work by Kuha and Skinner (1997). A relatively recent overview is provided by Buonaccorsi (2010).

The literature on misclassification bias shows that the bias can be reduced significantly, if some form of extra information is available. In the general context of categorical data analysis, this extra information can be, for instance, replicate values, validation data, or instrumental variables (Buonaccorsi, 2010). Although such extra information in general might not always be available, it is available in the context of supervised machine learning that we are considering here. The extra information are validation data, which are traditionally used for model selection, training and testing. We will use the test set as validation data to estimate error rates, and thus to correct misclassification bias.

In experimental projects at NSIs, the test set often is a random sample from the target population (e.g., all households in the country). The setup corresponds to the double sampling scheme introduced by Tenenbein (1970). Among the correction methods discussed by Buonaccorsi (2010), the so-called *calibration estimator* then outperforms all the others in terms of the MSE, as proved theoretically by Kloos, Meertens, Scholtus and Karch (2020).

However, a new problem arises when incorporating machine learning algorithms in the production process of official statistics. There, a statistical model is often estimated once and then applied for a longer period of time without updating the model parameters. In the context of supervised machine learning this is common, because otherwise new data have to be annotated manually in each time period leading to high production costs. However, the problem there is that the data distribution as well as the relation between the dependent and independent variables might change over time, causing the outcome of the model to be biased. In the machine learning literature, this problem is known as *concept drift*. It has been investigated in stream learning and online learning for several decades (see Widmer & Kubat, 1996), dating back at least to the work on incremental learning (cf. Schlimmer & Granger, 1986) in the 1980s. Originally, the term *concept* was used for a set of Boolean-valued functions (Helmbold & Long, 1994). Currently, it has a statistical interpretation that is more closely related to our setting. Nowadays, Webb et al. (2016) state that the term *concept* refers to the joint distribution $\mathbb{P}(Y, X)$, with class labels (dependent variable) Y and features (independent variables) X, as proposed by Gama et al. (2014). Allowing such a joint distribution to depend on a time parameter t, concept drift in the setting of supervised learning means that $\mathbb{P}_{t_1}(Y, X) \neq \mathbb{P}_{t_2}(Y, X)$, for $t_1 \neq t_2$. The effect of concept drift is that misclassification bias might increase even further.

In this paper, we aim to prove which of the two popular correction methods discussed by Buonaccorsi (2010) reduces the MSE of statistical output most, under a specific type of concept drift known as *prior probability shift* (Moreno-Torres et al., 2012). Our paper deliberately focuses on the production process (where concept drift arises), building on the results obtained by Kloos et al. (2020) for the preceding experimental phase. Our numerical analyses will show, for the first time, that a decision boundary arises. The optimal choice for a correction method depends on three parameters, viz. the class distribution (or class imbalance), the size of the test set, and the model accuracy. With that knowledge we aim to contribute to the literature on concept drift *understanding* as defined by Lu et al. (2019). It complements concept drift *quantification* (Goldenberg & Webb, 2019) and concept drift *adaptation* (Gama et al., 2014). Analysing the decision boundary as a function of the three parameters yields practical recommendations for the implementation of classification algorithms in the production process of official statistics. Finally, analysing the impact of the *size* of the (manually created) test set allows us to comment on the cost efficiency of official statistics based on classification algorithms.

The remainder of the paper is organised as follows. In Section 3.2, we provide expressions for the bias and variance of the misclassification and calibration estimator, when applied to machine learning algorithms that have been implemented in the production process of official statistics. We show (1) that the optimal correction method in the experimental phase is no longer unbiased when implemented in a production process and we provide (2) a sharp lower bound for the absolute value of its bias. Hence, instead of arriving at a conclusive optimal solution in the experimental phase, a decision boundary arises in the context of the production process. Subsequently, in Section 3.3, we investigate the location and shape of that decision boundary. In Section 3.4 we present our conclusions and suggest three promising directions for future research.

3.2 Methods

In the context of official statistics, the convention is to use the MSE to evaluate output quality, also when using statistical models (Buelens et al., 2016). The key question when correcting misclassification bias then becomes: which correction method reduces the MSE of the output most? The outcome depends on the assumptions made. The situation that fits the experimental phase of machine learning projects at NSIs is discussed briefly in Subsection 3.2.1. The assumptions made in the experimental phase are considered to be the most restrictive ones. The answer to the key question under those restrictive assumptions has been provided by Kloos et al. (2020) and it is rather conclusive. A drawback of their result is that data are assumed to be annotated manually in each time period. In practice, manual data annotation is time consuming and hence expensive. Therefore, in Subsection 3.2.2, we describe the situation that corresponds to the production process of official statistics. In Subsection 3.2.3, the theoretical results known for the experimental phase are adapted to suit the conditions of the production process of official statistics. The answer to the key question in that setting is presented in Section 3.3.

3.2.1 The experimental phase

Consider a population I of N objects (households, enterprises, aerial images, company websites or other text documents) and some target classification, or stratum, s_i for each object $i \in I$. For now, we restrict ourselves to dichotomous categorical variables, i.e., $s_i \in \{0, 1\}$, where category 1 indicates the category of interest. A compelling example is the use of aerial images of rooftops to identify houses (the objects indexed by i) with solar panels ($s_i = 1$) (Curier et al., 2018). From now on, we make three essential assumptions. Our first assumption is that there is some (possibly time consuming or otherwise expensive) way to retrieve the true category s_i for each $i \in I$, for example by manually inspecting the aerial images and annotating them with a label indicating whether the image contains a solar panel. Our second assumption is that background variables or other features in the data contain sufficient information to estimate s_i accurately. We draw a small random sample from the population and determine the true category s_i for the objects in the sample. Then, the obtained data are, as usual, split at random into two sets. The first set is used to estimate model parameters (model selection and training). The second set, referred to as the test set $I_{\text{test}} \subset I$, is used to estimate the out-of-sample prediction error of the model. The number of observations in the test set is denoted by n and we assume that $n \ll N$.

Consequently, the model can be used to produce an estimate $\widehat{s}_i$ of the true category to which object i belongs. Here, our third assumption is that the success and misclassification probabilities of the model depend on i, but only through the true value of s_i. More precisely, we let p_{ab} be the probability that $\widehat{s}_i = b$ given that $s_i = a$, for $a, b \in \{0, 1\}$. This specifies the *classification error model* as introduced by Bross (1954), following the notation in Van Delden et al. (2016). In addition, we adopt the notation $\boldsymbol{a}_i$, which is a 2-vector equal to $(1, 0)$ if $s_i = 1$ and $(0, 1)$ if $s_i = 0$.

The estimate $\widehat{\boldsymbol{a}}_i$ is defined similarly. The sum of all $\boldsymbol{a}_i$ is the 2-vector of counts $\boldsymbol{v}$. The first component of the 2-vector $\boldsymbol{\alpha} = \boldsymbol{v}/N$ is called the *base rate* and is denoted by α. It is immediate that $\mathbb{E}[\widehat{\boldsymbol{\alpha}}] = P^T\boldsymbol{\alpha}$, where P is the confusion matrix with entries p_{ab} (with p_{11} as the top left entry). In general, $P^T\boldsymbol{\alpha} \neq \boldsymbol{\alpha}$, which indicates that $\widehat{\alpha}$ is a biased estimator of the base rate α. The statistical bias of $\widehat{\alpha}$ as estimator of the base rate α is referred to as *misclassification bias*.

A wide range of correction methods to reduce misclassification bias is available, see Buonaccorsi (2010). As briefly indicated in Section 3.1, Kloos et al. (2020) compared several correction methods aimed at improving the accuracy of estimators for α. Two correction methods were most promising. The first correction method is the *misclassification estimator* $\widehat{\alpha}_p$. It is defined as the first component of the following 2-vector:

$$\widehat{\boldsymbol{\alpha}}_p = \left(\widehat{P}^T\right)^{-1}\widehat{\boldsymbol{\alpha}}, \tag{3.1}$$

in which $\widehat{P}$ is the row-normalised confusion matrix obtained from the test set, i.e., with entries $\widehat{p}_{ab} = n_{ab}/n_{a+}$, where n_{ab} denotes the number of objects i in the test set for which $s_i = a$ and $\widehat{s}_i = b$ and where n_{a+} denotes $n_{aa} + n_{ab}$. Moreover, the second correction method is the *calibration estimator* $\widehat{\alpha}_c$. It is defined as the first component of the following 2-vector:

$$\widehat{\boldsymbol{\alpha}}_c = \widehat{C}\widehat{\boldsymbol{\alpha}}, \tag{3.2}$$

in which $\widehat{C}$ is the column-normalised confusion matrix obtained from the test set, i.e., with entries $\widehat{c}_{ab} = n_{ab}/n_{+b}$, where n_{+b} denotes $n_{ab} + n_{bb}$. Kloos et al. (2020) have shown that if the test set is indeed a random sample from the target population, then the MSE of $\widehat{\alpha}_c$ is always smaller than that of $\widehat{\alpha}_p$.

3.2.2 The production process of official statistics

Official statistics on a particular social or economic indicator are often produced for a certain period of time, at least annually, but often more frequently (quarterly or monthly). For as long as NSIs produce the official statistics on such an indicator, the output quality is required to be high. A challenging element in using classification algorithms in the production process of official statistics is that the target population I changes over time, including the background variables $\boldsymbol{x}_i$ and the base rate α. Therefore, the test set drawn at random from the population at one time period cannot be viewed as a random sample from the population at the next time period. A first solution would be to draw a new test set from the population (and then manually annotate the data) at each time period for as long as the statistical indicator is produced. However, due to cost constraints, such frequent data annotation is infeasible in practice. Thus, we will have to make an additional assumption to further investigate the results achieved by Kloos et al. (2020) in the context of a production process.

The additional assumption that we make is that the out-of-sample prediction accuracy of the model, i.e., the matrix P, is stable during a short period of time. More specifically, we assume (1) that s_i causally determines the background variables $\boldsymbol{x}_i$ that are used in the model for $\widehat{s}_i$ and (2) that the causal relation does not change between (at least) two consecutive months or quarters. These two assumptions are identical to *prior probability shift* as defined by Moreno-Torres et al. (2012). The first assumption, i.e., the causal relation between s_i and $\boldsymbol{x}_i$, seems reasonable in many applications. In epidemiology, a disease causally determines the symptoms. In sentiment analysis, the writer's sentiment causally determines the words that the writer chooses. In land cover mapping, the mapped object causally determines the pixel values in the image. The second assumption (in terms of the classification error model) reads that $\mathbb{P}(\widehat{s}_i | s_i)$ does not change between consecutive months or quarters, but that α is allowed to change.

In the setting of prior probability shift, we consider two populations, namely the target population at two different moments in time, indicated by I and I', with sizes N and N'. We assume that the test set $I_{\text{test}} \subset I$ of size n has been obtained as a random sample from the target population I in the first month or quarter, with true base rate α. The aim is to estimate the base rate α' in the second month or quarter, i.e., within population I', using prediction $\widehat{s}_i$ for $i \in I'$ and the estimates of p_{ab} based on $I_{\text{test}} \subset I$. The type of concept drift that we investigate, prior probability shift, can be quantified by the difference $\delta := \alpha' - \alpha$, which we will briefly refer to as the *drift*. In the experimental phase we only consider a single population, which corresponds to putting $\delta = 0$. In Subsection 3.2.3, we investigate the MSE of the calibration and misclassification estimator when $\delta \neq 0$.

3.2.3 Theoretical results

Expressions for the bias and variance of the misclassification estimator α_p under drift δ can be derived easily from the expressions presented by Kloos et al. (2020).

It follows that

$$
\begin{aligned}
B[\hat{\alpha}_p] &= \frac{1}{n(p_{00}+p_{11}-1)^2}\cdot\left[\frac{\alpha'}{\alpha}p_{11}(1-p_{11}) - \frac{1-\alpha'}{1-\alpha}p_{00}(1-p_{00})\right] + O\left(\frac{1}{n^2}\right) \\
&= \frac{p_{00}-p_{11}}{n(p_{00}+p_{11}-1)} \\
&\quad + \frac{\delta}{n(p_{00}+p_{11}-1)^2}\cdot\left(\frac{p_{11}(1-p_{11})}{\alpha}+\frac{p_{00}(1-p_{00})}{1-\alpha}\right) + O\left(\frac{1}{n^2}\right),
\end{aligned}
\tag{3.3}
$$

which is increasing in δ (but might first decrease in δ in absolute value). The variance of the misclassification estimator equals

$$
V(\hat{\alpha}_p) = \frac{(1-\alpha')^2 V(\hat{p}_{00}) + \alpha'^2 V(\hat{p}_{11})}{(p_{00}+p_{11}-1)^2} + O\left(\frac{1}{n^2}\right),
\tag{3.4}
$$

We neglect the terms of order $1/n^2$ and use Equations (3.8) and (3.9) from Appendix 3.A to obtain

$$
\begin{aligned}
V(\hat{\alpha}_p) = \frac{1}{n(p_{00}+p_{11}-1)^2}\cdot\Bigg[&T + 2\delta(p_{00}-p_{11})(p_{00}+p_{11}-1) \\
&+ \delta^2\cdot\left(\frac{p_{11}(1-p_{11})}{\alpha}+\frac{p_{00}(1-p_{00})}{1-\alpha}\right)\Bigg] + O\left(\frac{1}{n^2}\right),
\end{aligned}
\tag{3.5}
$$

in which $T := (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$. If $p_{00} \geq p_{11}$, then the variance increases as the drift δ increases. If $p_{00} < p_{11}$, then the effect of the drift is not immediately clear: increasing δ might decrease the variance, depending on the values of α and δ. In Section 3.3, we will analyse the behaviour of $V(\hat{\alpha}_p)$ as function of α and δ numerically.

The expressions for the bias and variance of the calibration estimator presented by Kloos et al. (2020) were derived by conditioning on the base rate in the target population. If the drift δ is nonzero, that proof strategy breaks down. Therefore, we have adapted the proof to hold for nonzero δ, resulting in the following expressions (see (3.6) and (3.7)).

Theorem 3.1. *The bias of $\hat{\alpha}_c$ as estimator of α under drift δ is given by*

$$
B[\hat{\alpha}_c] = -\delta\frac{T}{\beta(1-\beta)} + O\left(\frac{1}{n^2}\right),
\tag{3.6}
$$

in which $\beta := (1-\alpha)(1-p_{00}) + \alpha p_{11}$ and $T = (1-\alpha)p_{00}(1-p_{00}) + \alpha p_{11}(1-p_{11})$. With that notation, the variance of $\hat{\alpha}_c$, under drift δ, is given by

$$V(\hat{\alpha}_c) = \frac{\alpha(1-\alpha)}{n}\left[\frac{T}{\beta(1-\beta)} + 2\delta(p_{00}+p_{11}-1)\left(\frac{p_{11}(1-p_{00})}{\beta^2} - \frac{p_{00}(1-p_{11})}{(1-\beta)^2}\right)\right.$$
$$\left. + \delta^2(p_{00}+p_{11}-1)^2\left(\frac{p_{11}(1-p_{00})}{\beta^3} + \frac{p_{00}(1-p_{11})}{(1-\beta)^3}\right)\right] + O\left(\frac{1}{n^2}\right). \tag{3.7}$$

Proof. See Appendix 3.A. □

We make the following two observations: (1) the bias and the drift δ have opposite signs. and (2) the absolute bias is linearly increasing as a function of the absolute drift $|\delta|$. From these observations, the following sharp upper bound and lower bound for the absolute bias in terms of the absolute drift can be derived.

Theorem 3.2. *The absolute bias of $\hat{\alpha}_c$ as estimator of $\alpha' = \alpha + \delta$ is bounded from above by $|\delta|$. If $p_{00} \leq p$ and $p_{11} \leq p$ for some $1/2 \leq p \leq 1$, then the absolute bias is at least $4p(1-p)|\delta|$.*

Proof. See Appendix 3.A. □

The third observation is that, under prior probability shift, the bias of the misclassification estimator is still of order $1/n$ while that of the calibration estimator is nonzero if $\delta \neq 0$ and does not decrease for increasing n. This third observation is the key observation. The implication is that the conclusions drawn by Kloos et al. (2020) for the experimental phase of a machine learning project in official statistics do not hold when the algorithms are implemented in the production process. There, the drift δ is nonzero and a decision boundary arises. The aim of Section 3.3 is to investigate the properties of the decision boundary.

3.3 Results

The theoretical results from Section 3.2 indicate that in case δ is nonzero a decision boundary arises (between preferring (a) the misclassification estimator and (b) the calibration to reduce misclassification bias). The aim of this section is to understand that decision boundary. It is the main focus of Subsection 3.3.3. In advance, we investigate the bias under prior probability shift of the calibration

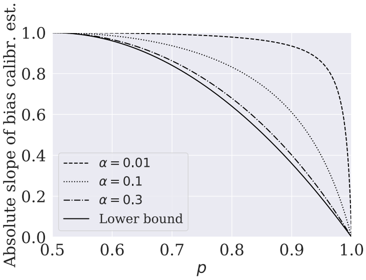


FIG. 3.1: The slope of the bias of the calibration estimator $\hat{\alpha}_c$ as a function of the drift δ is equal to $-T/(\beta(1-\beta))$, which is strictly negative. The absolute value of that slope is plotted against the classification probability p, assuming that $p_{00} = p_{11} = p$, for four different values of α. The solid black line depicts the theoretical lower bound (see Theorem 3.2) for the slope of the bias.

estimator more closely in Subsection 3.3.1 and the difference in MSE between the two estimators in Subsection 3.3.2.

3.3.1 Bias of the calibration estimator

We start plotting $T/(\beta(1-\beta))$, the absolute value of the slope of the bias of the calibration estimator, as a function of the classification probabilities for different values of α, i.e., the base rate in the test set. For visualisation purposes, we restrict the function to $p_{00} = p_{11}$, parameterised by p. The results are depicted in Fig. 3.1, including the theoretical lower bound stated in Theorem 3.2. The slope of the bias as a function of p is decreasing from 1 at $p = 0.5$ to 0 at $p = 1$. The smaller the value of α, the later the function drops to 0. The reason is that the drift δ is defined as an absolute number and therefore it is relatively larger for smaller values of α. From this observation we may conclude that the impact of (an absolute) drift δ on the bias of $\hat{\alpha}_c$ increases if α is further away from 0.5, i.e., if the so-called *class imbalance* increases.

3.3.2 Difference in mean squared error

Subsequently, we investigate the difference $D(\hat{\alpha}_p, \hat{\alpha}_c) := MSE(\hat{\alpha}_p) - MSE(\hat{\alpha}_c)$ between the MSE of the misclassification estimator and that of the calibration estimator. The value of $D(\hat{\alpha}_p, \hat{\alpha}_c)$ as a function of δ is depicted in Fig. 3.2 for each

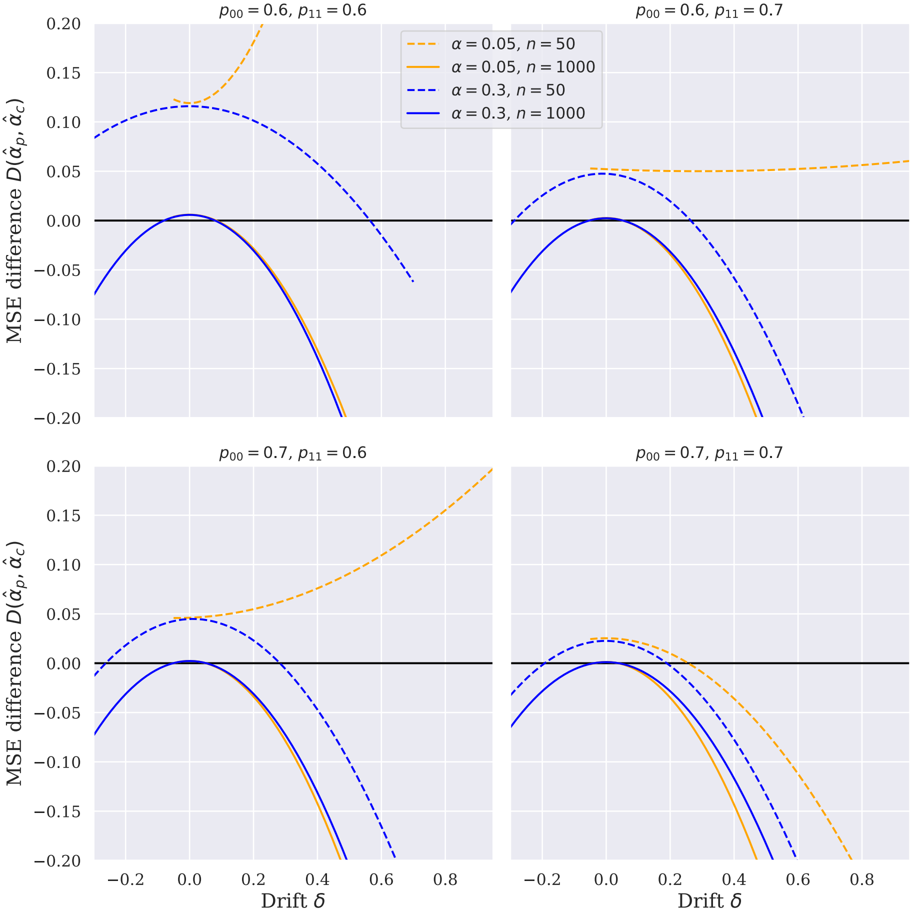


Fig. 3.2: The difference $D(\hat{\alpha}_p, \hat{\alpha}_c)$ between the MSE of the misclassification estimator $\hat{\alpha}_p$ and that of the calibration estimator $\hat{\alpha}_c$, plotted as a function of δ for each possible combination of $\alpha \in \{0.05, 0.3\}$, $n \in \{50, 1000\}$ and $p_{00}, p_{11} \in \{0.6, 0.7\}$. Note that the drift δ ranges from $-\alpha$ to $1 - \alpha$, because $\alpha' = \alpha + \delta$ must lie between 0 and 1.

possible combination of $\alpha \in \{0.05, 0.3\}$, $n \in \{50, 1000\}$ and $p_{00}, p_{11} \in \{0.6, 0.7\}$. Note that the drift δ ranges from $-\alpha$ to $1-\alpha$, because $\alpha' = \alpha+\delta$ must lie between 0 and 1. We report the following four observations. First, the difference is positive if $\delta = 0$ in any of the line plots, which corresponds to the main conclusion drawn by Kloos et al. (2020). Second, when n is sufficiently large (thin lines), the difference between the line plots are small. The reason is that the contribution of the variance terms is negligible compared to that of the squared bias of $\hat{\alpha}_c$, which does not depend on n (see Theorem 3.1). Third, for highly imbalanced data sets combined with small test sets, i.e., α close to 0 and n small (thick dash-dotted lines), the variance of $\hat{\alpha}_p$ dominates if either p_{00} is close to 0.5 or p_{11} is close to 0.5. As a result, the calibration estimator has the smallest MSE, independent of the magnitude of the drift δ. Fourth, if the class distribution is relatively balanced (dotted lines), the difference $D(\hat{\alpha}_p, \hat{\alpha}_c)$ will become negative if δ increases, but the intersection moves farther away from $\delta = 0$ as n decreases.

3.3.3 The preferred estimator

Finally, we compute, numerically, the unique positive value of δ (if it exists) at which the MSE of the misclassification and calibration estimator are identical. That is, we collect and reorganise the points of intersection $D(\hat{\alpha}_p, \hat{\alpha}_c) = 0$ as discussed in Subsection 3.3.2. We view $D(\hat{\alpha}_p, \hat{\alpha}_c)$ as a map from $\mathbb{R}^3$ to $\mathbb{R}$ by fixing α and n and using δ, p_{00} and p_{11} as variables. Then, we plot the line within the two-dimensional surface $D(\hat{\alpha}_p, \hat{\alpha}_c) = 0$ where $p_{00} = p_{11}$, resulting in Fig. 3.3. Interestingly, the result is a decreasing function of p. At first, the result might seem to contradict the result obtained in the first analysis, cf. Fig. 3.1. There, the absolute slope of the bias as function of δ decreases with increasing p. Hence, the MSE of $\hat{\alpha}_c$ increases more slowly as a function of δ with increasing p. However, the result in Fig. 3.3 follows from the fact that the difference in variance between $\hat{\alpha}_c$ and $\hat{\alpha}_p$ rapidly decreases as p increases.

We stress that the lines in Fig. 3.3 can be interpreted as decision boundaries. Each statistical indicator that is based on a classification algorithm plots somewhere in the (p, δ)-plane depicted in Fig. 3.3. Our experimental result then reads as follows. If the plot of the indicator in the (p, δ)-plane ends up above the decision boundary (which depends on α and n), then the misclassification estimator should be preferred over the calibration estimator to reduce misclassification bias. Otherwise, the calibration estimator should be preferred over the misclassification

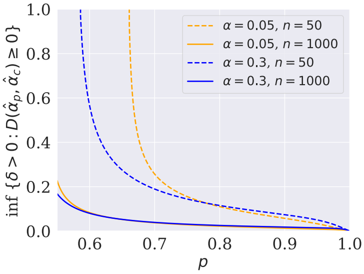


FIG. 3.3: The unique positive value δ (if it exists) for which $D(\hat{\alpha}_p, \hat{\alpha}_c) = 0$, as a function of the classification probability p, assuming $p_{00} = p_{11} = p$. The lines should be interpreted as decision boundaries: below each of these lines the calibration estimator is preferred, while above each of the lines the misclassification estimator is preferred.

estimator. Moreover, in practice one should always compute the (estimated) bias and variance of the applied estimator, for they might still be large, e.g., when n and p are small and δ is large.

As a final remark, we indicate that these results hold if only the misclassification estimator and calibration estimator are considered. Admittedly, there may exist other estimators that might reduce misclassification bias even further.

3.4 Chapter conclusions

In this research, we investigated the output quality of official statistics based on classification algorithms. The main problem examined was how to reduce the bias caused by prior probability shift. We focused on two bias correction methods, namely (1) the misclassification estimator and (2) the calibration estimator. The results known for these two estimators failed to hold under prior probability shift. To obtain a further insight into the output quality of official statistics based on classification algorithms under prior probability shift, we adapted and extended the results achieved by Kloos et al. (2020) to hold for any value of the drift δ. As theoretical results, we were able to show that (1) the calibration estimator is no longer unbiased and that (2) the absolute bias as a first-order approximation is a linearly increasing function of the absolute drift $|\delta|$ and does not depend on the test set size n.

Building on the theoretical results, we performed a simulation study consisting of three subsequent numerical analyses. The main conclusion drawn from the simulation results, is that the MSE of the calibration estimator is smaller than that of the misclassification estimator only when the performance of the classifier (in terms of p_{00} and p_{11}) is low or when the drift δ is close to 0. The main conclusion has at least two significant implications. The first implication is that the conclusion gives a better understanding of the output quality of official statistics based on machine learning algorithms. More specifically, recommendations on which correction methods should be implemented in which situation are given. They allow for a more reliable implementation of machine learning algorithms in official statistics. The second implication is that the impact of the size and frequency of the training and test data sets is better understood. Essentially, our results show that the calibration estimator should *not* be applied to data streams or time series data, unless training and test data in each time period are available to (a) retrain the classifier and hence (b) adapt to concept drift.

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

Finally, we suggest three directions for future research. First, the robustness of classifier-based estimators should also be investigated for other types of concept drift, starting with the less restrictive type of prior probability shift as defined by Webb et al. (2016). Second, it might be worthwhile to examine methods for concept drift adaptation that are based on unlabelled data only, by carefully incorporating changes in the distribution of $P(X)$. Third, combinations or ensembles of different estimators require further research. We believe that a well-chosen combination of estimators will increase the overall robustness of classifier-based estimators under concept drift.

APPENDIX

3.A Theoretical derivations under assumption A2

This appendix contains the proofs of the theorems presented in Chapter 3.6. For clarity, we will write $\widehat{\alpha}^*$ for the estimator based on the algorithms predictions $\widehat{s}_i$, see also Chapter 2. In addition to the assumptions described in Section 3.2, we make two more technical assumptions, namely that $\widehat{\alpha}^*$ is independent of both the $\hat{c}_{ij}$ and the $\hat{p}_{ij}$. It follows that $\hat{p}_{00}$ and $\hat{p}_{11}$ are uncorrelated and that

$$V(\hat{p}_{11}) = \frac{p_{11}(1-p_{11})}{n\alpha}\left[1 + \frac{1-\alpha}{n\alpha}\right] + O\left(\frac{1}{n^3}\right). \tag{3.8}$$

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

$$V(\hat{p}_{00}) = \frac{p_{00}(1-p_{00})}{n(1-\alpha)}\left[1 + \frac{\alpha}{n(1-\alpha)}\right] + O\left(\frac{1}{n^3}\right). \tag{3.9}$$

For the proofs of these statements, consult Lemma 1 in Appendix 2.A. We will now provide the proof of Theorem 3.1 below.

Proof of Theorem 3.1. Recall that the calibration estimator $\hat{\alpha}_c$ was given by

$$\hat{\alpha}_c = \hat{\alpha}^*\hat{c}_{11} + (1 - \hat{\alpha}^*)\hat{c}_{10}. \tag{3.10}$$

The derivations of the bias $B[\hat{\alpha}_c]$ and $V[\hat{\alpha}_c]$ are included below.

Bias. It is assumed that $\hat{\alpha}^*$ and $\hat{c}_{ij}$ are independent. Hence,

$$\mathbb{E}[\hat{\alpha}_c] = \mathbb{E}[\hat{\alpha}^*]\,\mathbb{E}[\hat{c}_{11}] + \mathbb{E}[1 - \hat{\alpha}^*]\,\mathbb{E}[\hat{c}_{10}]. \tag{3.11}$$

Recall the notation $\beta = (1-\alpha)(1-p_{00}) + \alpha p_{11}$ and set $\beta' := (1-\alpha')(1-p_{00}) + \alpha' p_{11} = \mathbb{E}[\hat{\alpha}^*]$. To compute $\mathbb{E}[\hat{c}_{ij}]$, condition on n_{1+}, and note that $n_{0+} = n - n_{1+}$ is n_{1+}-measurable. It holds that $\hat{c}_{11} \mid n_{1+} \overset{d}{=} X/(X+Y)$, with $X \sim Bin(n_{1+}, p_{11})$ and $Y \sim Bin(n_{0+}, 1-p_{00})$. Introducing the stochastic variable $\beta_+ := n_{1+}p_{11} + n_{0+}(1-p_{00})$, a second-order Taylor approximation yields

$$\begin{aligned}\mathbb{E}[\hat{c}_{11} \mid n_{1+}] &= \frac{n_{1+}p_{11}}{\beta_+} - \frac{n_{0+}(1-p_{00})}{\beta_+^3} n_{1+}p_{11}(1-p_{11}) \\ &\qquad + \frac{n_{1+}p_{11}}{\beta_+^3} n_{0+}p_{00}(1-p_{00}) + O\left(\frac{1}{n^2}\right) \\ &= \frac{n_{1+}p_{11}}{\beta_+} + p_{11}(1-p_{00})(p_{00}+p_{11}-1)\frac{n_{0+}n_{1+}}{\beta_+^3} + O\left(\frac{1}{n^2}\right). \qquad (3.12)\end{aligned}$$

We then introduce the random variable $Z \sim Bin(n, \alpha)$ (i.e., $Z \overset{d}{=} n_{1+}$). Applying a Taylor approximation to the first term of expression (3.12) yields

$$\begin{aligned}\mathbb{E}\left[\frac{n_{1+}p_{11}}{\beta_+}\right] &= \mathbb{E}\left[\frac{p_{11}Z}{n(1-p_{00}) + (p_{00}+p_{11}-1)Z}\right] \\ &= \frac{\alpha p_{11}}{\beta} - \frac{1}{2}\frac{2np_{11}(1-p_{00})(p_{00}+p_{11}-1)}{n^3\beta^3} n\alpha(1-\alpha) + O\left(\frac{1}{n^2}\right) \\ &= c_{11} - \frac{\alpha(1-\alpha)}{n}\frac{p_{11}(1-p_{00})(p_{00}+p_{11}-1)}{\beta^3} + O\left(\frac{1}{n^2}\right). \qquad (3.13)\end{aligned}$$

Next, apply a Taylor approximation to (the stochastic part of) the second term in expression (3.12):

$$\mathbb{E}\left[\frac{Z(n-Z)}{\beta_+^3}\right] = \frac{\alpha(1-\alpha)}{n\beta^3} + O\left(\frac{1}{n^2}\right). \qquad (3.14)$$

Combining equations (3.13) and (3.14) results in

$$\mathbb{E}[\hat{c}_{11}] = c_{11} + O\left(\frac{1}{n^2}\right) = \frac{\alpha p_{11}}{\beta} + O\left(\frac{1}{n^2}\right), \qquad (3.15)$$

where the second equality is included to stress that the result depends on α, and not on α'. Similarly, it follows that

$$\mathbb{E}[\hat{c}_{10}] = c_{10} + O\left(\frac{1}{n^2}\right) = \frac{\alpha(1-p_{11})}{1-\beta} + O\left(\frac{1}{n^2}\right). \tag{3.16}$$

Substituting $\alpha' = \alpha + \delta$ and neglecting terms of order $1/n^2$ yields

$$\begin{aligned}\mathbb{E}[\hat{\alpha}_c] &= \beta'\frac{\alpha p_{11}}{\beta} + (1-\beta')\frac{\alpha(1-p_{11})}{1-\beta}\\ &= \alpha p_{11} + \delta(p_{00}+p_{11}-1)\frac{\alpha p_{11}}{\beta} + \alpha(1-p_{11}) + \delta(1-p_{00}-p_{11})\frac{\alpha(1-p_{11})}{1-\beta}\\ &= \alpha + \frac{\delta\alpha}{\beta(1-\beta)}\Big((1-\beta)p_{11} - \beta(1-p_{11})\Big)(p_{00}+p_{11}-1)\\ &= \alpha + \frac{\delta\alpha(1-\alpha)(p_{00}+p_{11}-1)^2}{\beta(1-\beta)}. \end{aligned} \tag{3.17}$$

It is straightforward to check that

$$\beta(1-\beta) - \alpha(1-\alpha)(p_{00}+p_{11}-1)^2 = \alpha p_{11}(1-p_{11}) + (1-\alpha)p_{00}(1-p_{00}) =: T. \tag{3.18}$$

Hence,

$$\mathbb{E}[\hat{\alpha}_c] = \alpha + \delta\left(\frac{\beta(1-\beta)-T}{\beta(1-\beta)}\right) + O\left(\frac{1}{n^2}\right) = \alpha' - \delta\frac{T}{\beta(1-\beta)} + O\left(\frac{1}{n^2}\right). \tag{3.19}$$

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

$$B[\hat{\alpha}_c] = -\delta\frac{T}{\beta(1-\beta)} + O\left(\frac{1}{n^2}\right). \tag{3.20}$$

Variance. To compute the variance of $\hat{\alpha}_c$, we first note that

$$\mathbb{E}[(\hat{\alpha}^*)^2] = \mathbb{E}[\hat{\alpha}^*]^2 + V(\hat{\alpha}^*) = \mathbb{E}[\hat{\alpha}^*]^2 + O\left(\frac{1}{N}\right). \tag{3.21}$$

A similar expression holds for the expectation of $(1-\hat{\alpha}^*)^2$ and that of $(1-\hat{\alpha}^*)\hat{\alpha}^*$. Neglecting the terms of order $1/N$, the above implies that

$$\begin{aligned} V(\hat{\alpha}_c) &= V(\hat{\alpha}^*\hat{c}_{11}) + V((1-\hat{\alpha}^*)\hat{c}_{10}) + C(\hat{\alpha}^*\hat{c}_{11},(1-\hat{\alpha}^*)\hat{c}_{10}) \\ &= \mathbb{E}[\hat{\alpha}^*]^2 V(\hat{c}_{11}) + \mathbb{E}[1-\hat{\alpha}^*]^2 V(\hat{c}_{10}) + \mathbb{E}[\hat{\alpha}^*]\,\mathbb{E}[(1-\hat{\alpha}^*)]C(\hat{c}_{11},\hat{c}_{10}). \end{aligned} \tag{3.22}$$

We may already substitute $\mathbb{E}[\hat{\alpha}^*] = \beta'$ in the above. It remains to derive expressions for $V(\hat{c}_{11})$, $V(\hat{c}_{10})$ and $C(\hat{c}_{11},\hat{c}_{10})$. We compute $V(\hat{c}_{11})$ as $\mathbb{E}[\hat{c}_{11}^2] - \mathbb{E}[\hat{c}_{11}]^2$, because we have already derived an expression for the latter term. The random variable $\hat{c}_{11}^2 \mid n_{1+}$ is distributed as $X^2/(X+Y)^2$. Setting $f(x,y) = x^2/(x+y)^2$ yields

$$f_{xx}(x,y) = \frac{2y^2 - 4xy}{(x+y)^4}, \quad \text{and} \quad f_{yy}(x,y) = \frac{6x^2}{(x+y)^4}. \tag{3.23}$$

It follows, neglecting terms of higher order, that

$$\begin{aligned} &\mathbb{E}[\hat{c}_{11}^2 \mid n_{1+}] \\ &= \frac{n_{1+}^2 p_{11}^2}{\beta_+^2} + \frac{n_{0+}^2(1-p_{00})^2 - 2n_{1+}n_{0+}p_{11}(1-p_{00})}{\beta_+^4} n_{1+}p_{11}(1-p_{11}) \\ &\qquad + \frac{3n_{1+}^2 p_{11}^2}{\beta_+^4} n_{0+}p_{00}(1-p_{00}) \\ &= \frac{n_{1+}^2 p_{11}^2}{\beta_+^2} + p_{11}(1-p_{00}) \frac{n_{1+}n_{0+}\Big(n(1-p_{00})(1-p_{11}) + n_{1+}(p_{00}+p_{11}-1)(2p_{11}+1)\Big)}{\beta_+^4}. \end{aligned} \tag{3.24}$$

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

$$f_{zz}(z) = \frac{2A^2 - 4ABz}{(A+Bz)^4}. \tag{3.25}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\frac{n_{1+}^2 p_{11}^2}{\beta_+^2}\right] \\
&\quad = \mathbb{E}\left[\frac{p_{11}^2 Z^2}{(A+BZ)^2}\right] \\
&\quad = \frac{\alpha^2 p_{11}^2}{\beta^2} + p_{11}^2 \frac{n^2(1-p_{00})^2 - 2n^2\alpha(1-p_{00})(p_{00}+p_{11}-1)}{n^4\beta^4} n\alpha(1-\alpha) + O\left(\frac{1}{n^2}\right) \\
&\quad = c_{11}^2 + \frac{\alpha(1-\alpha)}{n} \frac{p_{11}^2(1-p_{00})\Big(1-p_{00}-2\alpha(p_{00}+p_{11}-1)\Big)}{\beta^4} + O\left(\frac{1}{n^2}\right). \qquad (3.26)
\end{aligned}$$

Apply a Taylor approximation to (the stochastic part of) the second term of expression (3.24) to obtain:

$$\frac{\alpha(1-\alpha)}{n} \frac{p_{11}(1-p_{00})\Big((1-p_{00})(1-p_{11}) + \alpha(p_{00}+p_{11}-1)(2p_{11}+1)\Big)}{\beta^4} + O\left(\frac{1}{n^2}\right). \qquad (3.27)$$

At last, combining expressions (3.26) and (3.27), and subtracting expression (3.15) squared, the variance of $\hat{c}_{11}$ can be expressed as

$$V(\hat{c}_{11}) = \frac{\alpha(1-\alpha)}{n} \frac{p_{11}(1-p_{00})}{\beta^3} + O\left(\frac{1}{n^2}\right). \qquad (3.28)$$

Similarly, it can be shown that

$$V(\hat{c}_{10}) = \frac{\alpha(1-\alpha)}{n} \frac{p_{00}(1-p_{11})}{(1-\beta)^3} + O\left(\frac{1}{n^2}\right). \qquad (3.29)$$

Moreover, it can be shown that $\hat{c}_{11}$ and $\hat{c}_{10}$ are uncorrelated. We use the same strategy that was used to prove that $\hat{p}_{00}$ and $\hat{p}_{11}$ are uncorrelated and we find that

$$\begin{aligned}\mathbb{E}[\hat{c}_{11}\hat{c}_{10}] &= \mathbb{E}\left[\mathbb{E}\left[\frac{n_{11}n_{10}}{n_{+1}n_{+0}}\middle| n_{+1}\right]\right] \\ &= \mathbb{E}\left[\frac{1}{n_{1+}n_{0+}}\,\mathbb{E}\left[n_{11}n_{10}\middle| n_{+1}\right]\right] \\ &= \mathbb{E}\left[\frac{1}{n_{1+}n_{0+}}\cdot n_{+1}c_{11}n_{+0}c_{10}\right] = c_{11}c_{10} = \mathbb{E}[\hat{c}_{11}]\,\mathbb{E}[\hat{c}_{10}]. \end{aligned} \tag{3.30}$$

It implies that $C(\hat{c}_{11}, \hat{c}_{10}) = \mathbb{E}[\hat{c}_{11}\hat{c}_{10}] - \mathbb{E}[\hat{c}_{11}]\,\mathbb{E}[\hat{c}_{10}] = 0$. Finally, we may conclude that

$$V(\hat{\alpha}_c) = \frac{\alpha(1-\alpha)}{n}\left(\beta'^2\frac{p_{11}(1-p_{00})}{\beta^3} + (1-\beta')^2\frac{p_{00}(1-p_{11})}{(1-\beta)^3}\right) + O\left(\frac{1}{n^2}\right). \tag{3.31}$$

3

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

$$\begin{aligned}V(\hat{\alpha}_c) = \frac{\alpha(1-\alpha)}{n}\Bigg[&\frac{T}{\beta(1-\beta)} + 2\delta(p_{00}+p_{11}-1)\left(\frac{p_{11}(1-p_{00})}{\beta^2} - \frac{p_{00}(1-p_{11})}{(1-\beta)^2}\right) \\ &+ \delta^2(p_{00}+p_{11}-1)^2\left(\frac{p_{11}(1-p_{00})}{\beta^3} + \frac{p_{00}(1-p_{11})}{(1-\beta)^3}\right)\Bigg] + O\left(\frac{1}{n^2}\right).\end{aligned} \tag{3.32}$$

The expression above completes the derivation of the variance of the calibration estimator under prior probability shift. □

To prove the Theorem 3.2, we need the following lemma.

Lemma 3.1. *The slope of the absolute value of the first-order approximation of the bias of the calibration estimator as a function of the absolute value* $|\delta|$ *of the prior probability shift is decreasing in* p_{00} *and* p_{11} *for all* $1/2 \leq p_{00} \leq 1$ *and* $1/2 \leq p_{11} \leq 1$.

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

$$f(x, y, \alpha) = \frac{(1-x)y}{\beta} \quad \text{and} \quad g(x, y, \alpha) = \frac{x(1-y)}{1-\beta}. \tag{3.33}$$

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

$$f_x(x, y, \alpha) = \frac{-\alpha y^2}{\beta^2} \quad \text{and} \quad g_x(x, y, \alpha) = \frac{\alpha(1-y)^2}{(1-\beta)^2}. \tag{3.34}$$

Hence,

$$h_x(x, y, \alpha) = \frac{\alpha}{\beta^2(1-\beta)^2} \cdot \Big(((1-y)\beta)^2 - (y(1-\beta))^2\Big). \tag{3.35}$$

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

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

We conclude that the slope h of the first-order approximation of the bias of the calibration estimator under prior probability shift is decreasing in p_{00} and p_{11} for $1/2 \le p_{00}, p_{11} \le 1$, attaining its global maximum at $p_{00} = p_{11} = 1/2$, where $h = 1$ and $\left|B[\hat{\alpha}_c]\right| = |\delta|$. □

Theorem 3.2 is an immediate consequence of the lemma above.

Proof of Theorem 3.2. Lemma 3.1 implies the two inequalities $|B[\hat{\alpha}_c]| \le |\delta|$ and $|B[\hat{\alpha}_c]| \ge |\delta| \cdot h(p, p, \alpha)$. To simplify the latter, observe that $T(p, p, \alpha) = p(1-p)$ and that $0 \le 1 - p < \beta(p, p, \alpha) < p \le 1$, using that $1/2 \le p \le 1$ and $0 < \alpha < 1$. It follows that $\beta(1-\beta) \le 1/4$, which completes the proof. □

CHAPTER 4

IMPOSING PARAMETER CONSTRAINTS

4.1 Introduction

Aggregation after automated classification naturally occurs in a wide range of data mining applications. Classifier-based aggregation even is the most common data operation in some research fields. Therefore, comprehending the effect of classification errors on the accuracy of the resulting aggregates is essential. The aim of this paper is to improve the accuracy of such aggregates by reducing their statistical bias. Before discussing technical details, we briefly describe two applications, demonstrating the relevance of the problem.

The first application is sentiment analysis in social media (Daas et al., 2015; Ravi & Ravi, 2015). For the sake of simplicity, assume that messages are either positive or negative and that the overall sentiment is defined as the difference between the number of positive and negative messages. The sentiment of one message is predicted using natural language processing. The estimator of the sentiment on social media obtained in this manner is statistically biased, unless precision equals recall, as we show below.

The second application is land cover mapping based on satellite imagery (Costa, Almeida, Vala, Marcelino & Caetano, 2018; L. Ma et al., 2017). Here, the aim is to estimate the area of different types of land cover (e.g., cropland, wetland) of a large, delimited surface (e.g, a country, a continent). In the paper by Costa et al. (2018), the per-pixel land cover class (one of 15) is predicted by an SVM image classifier. Again, the estimated surface per land cover class is biased.

The two applications above are examples of the following general setting. Consider a set of N data points. Each data point is equipped with a categorical variable s (for *stratum*) and a numerical variable y. We are interested in the aggregates obtained by summing y after grouping by s. We denote the resulting aggregates by the K-vector $\boldsymbol{u}$, where $K > 1$ is the number of categories that s may attain. Now, if s is not observed, but the result of a classification algorithm is used instead, classification errors emerge. We therefore distinguish between the true class s and the predicted class $\widehat{s}$. Similarly, we write $\widehat{\boldsymbol{u}}$ for the K-vector of *classifier-based statistics* obtained by summing y after grouping by $\widehat{s}$.

The problem studied in this paper is that the vector of classifier-based statistics $\widehat{\boldsymbol{u}}$ will be a *statistically biased* estimator of the true aggregate vector $\boldsymbol{u}$. The bias can be nonzero even if the accuracy of the classification algorithm is high and N is large. In fact, the example below (see the box titled "The Base Rate Example") shows that the bias does not depend on N at all.

The Base Rate Example

We consider a set of $N = 100{,}000$ companies, in which we would like to identify webshops based on the text found on the company's website. Assume a trained classification algorithm with false negative rate $p = p_{10} = \mathrm{FN}/(\mathrm{TP}+\mathrm{FN}) = 0.01$ and false positive rate $q = p_{01} = \mathrm{FP}/(\mathrm{TN}+\mathrm{FP}) = 0.005$. Assume that the set contains $v_1 = 10{,}000$ webshops, which would in practice be unknown. The fraction $v_1/N = 0.1$ is referred to as the *base rate*. The expected number of companies classified as webshops is

$$0.99 \cdot 10{,}000 + 0.005 \cdot 90{,}000 = 10{,}350, \tag{4.1}$$

showing that the estimator has a relative bias of +3.5%.

Essentially, The Base Rate Example shows that if s is binary and $y \equiv 1$,

$$\mathbb{E}[\widehat{\boldsymbol{u}}] = P^T \boldsymbol{u} \quad \text{with} \quad P = \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}. \tag{4.2}$$

4

In fact, it can be shown that Equation (4.2) holds for a multi-class variable s and any numerical variable y, with P being the $K \times K$ (row-normalised) *confusion matrix* (Van Delden et al., 2016). We now make two crucial observations. First, the *relative bias* $\mathbb{E}[(\widehat{\boldsymbol{u}} - u)/N]$ does *not* depend on N. It implies that the bias does not vanish for large data sets. Second, the bias is only equal to 0 if $p = q = 0$, or if (for constant y) the base rate is precisely equal to $q/(p+q)$. In the case of binary classification and constant y, the latter is equivalent to precision being equal to recall. In the case of multi-class classification and constant y, it is equivalent to the base rate vector (and hence $\boldsymbol{u}$) being an eigenvector (of the transposed confusion matrix P^T) corresponding to the eigenvalue 1.

Now, if the inverse $Q = (P^T)^{-1}$ of P^T is well-defined (in the binary case: if $p + q \neq 1$), then $Q\widehat{\boldsymbol{u}}$ is an unbiased estimator of $\boldsymbol{u}$. The problem is that the classification error rates are not known exactly and merely estimated. If the estimates of the classification error rates are based on a (very) small test set, the proposed unbiased estimator $Q\widehat{\boldsymbol{u}}$ might attain impermissible values. To illustrate this problem, we include a second example below.

The Peculiar Example

Consider a set of $N = 100$ companies in which the *predicted* number of webshops is equal to 10. To estimate the classification errors, a (very) small test set of size $n = 10$ is used. Assume that it resulted in TP = 4, FN = 1, FP = 2 and TN = 3, hence $p = 0.2$ and $q = 0.4$. Correcting the bias as suggested above yields

$$\left(P^T\right)^{-1} \widehat{\boldsymbol{u}} = \begin{pmatrix} 1.5 & -1 \\ -0.5 & 2 \end{pmatrix} \begin{pmatrix} 10 \\ 90 \end{pmatrix} = \begin{pmatrix} -75 \\ 175 \end{pmatrix}. \tag{4.3}$$

Thus, the unbiased estimate of the number of webshops in the data set is −75.

The issue in The Peculiar Example is caused by the fact that (1) p and q are not known but merely estimated and (2) the base rate is relatively low. Observe that the outcome does not dependent on the size N of the *full* data set. The problem arises because the *test* set is small (because it implies inaccurate estimates for p and q), as we will show in Section 4.3.2. Having only a small *test* set available, even if N is large, is quite common. The reason is that labelled data are unavailable in many applications (e.g., the sentiment analysis and land cover mapping examples), while manually creating labelled data requires expert knowledge, making it expensive.

We propose a novel bias correction method that reduces the statistical bias of classifier-based statistics. In contrast to existing methods, it is suitable for applications where the test set is small. If the test set *is* sufficiently large, our bias correction method will give as accurate results as existing methods.

This paper is structured as follows. In Section 4.2, the formal problem statement is introduced and related work is discussed. In Section 4.3, we formulate our bias correction method in the general setting of multi-class classification problems. In Section 4.4, we illustrate the effectiveness of the proposed methods using experiments on real-world data. Finally, Section 4.5 concludes.

4.2 Problem statement and related work

In this section, we introduce the notation and formal problem statement, and discuss related work.

4.2.1 Problem statement

We partly adopt the notation from Van Delden et al. (2016) to formulate our problem statement. Consider a set I of objects $i = 1, 2, \ldots, N$. Each object $i \in I$ belongs to a class s_i. The finite, ordered set of classes is denoted by H and is of size $K := |H| \geq 2$. In addition, each object is attributed with a continuous variable y_i of interest. We introduce the *class matrix* $A = (a_{ih})$ of dimension $N \times K$ given by $a_{ih} = \mathbb{I}(s_i = h)$, where $\mathbb{I}(\cdot)$ denotes the indicator function. The *counts vector* $\boldsymbol{v} := A^T\mathbf{1}$, where $\mathbf{1}$ is an N-vector of ones, counts the number of occurrences of the classes $h \in H$ in the population. That is, v_h equals the number of $i \in I$ for which $s_i = h$. The *base rate vector* $\boldsymbol{\alpha}$ of length K is given by $\alpha_h = v_h/N$ for $h \in H$.

Assume that the classes s_i are not known, but instead predicted to be $\widehat{s}_i$. The predicted class matrix based on the predictions $\widehat{s_i}$ is denoted by $\widehat{A}$. Similarly, $\widehat{\boldsymbol{v}}$ denotes the predicted counts vector. Assume that y_i is known for all $i \in I$. The goal is to estimate the sum of y_i over each of the classes $h \in H$. In other words, the main problem statement is how to find, using the predicted $\widehat{s}_i$ and the known y_i, an accurate estimator of the *aggregate vector*

$$\boldsymbol{u} = A^T y. \tag{4.4}$$

In the sentiment analysis application from before, the set I is the set of messages and s_i is binary (a positive or negative message). Each y_i equals 1 and hence $\boldsymbol{u}$ is a 2-vector with u_1 the number of positive messages and u_2 the number of negative messages. In the land cover mapping application obtained from the paper by Costa et al. (2018), each $i \in I$ is a pixel and s_i may attain $K = 15$ different values. The value y_i is the real land area corresponding to pixel i. Hence, the 15-vector $\boldsymbol{u}$ contains the total land area of each of the 15 land cover classes.

Now, a first estimator of $\boldsymbol{u}$ might be the K-vector

$$\widehat{\boldsymbol{u}} = \widehat{A}^T y. \tag{4.5}$$

However, we know that $\widehat{\boldsymbol{u}}$ is a biased estimator of $\boldsymbol{u}$, recall Equation (4.2). To estimate (and then correct) this statistical bias, we assume a *classification error model*, following the methodology of Van Delden et al. (2016): the value of $\widehat{s}_i$, given the value of s_i, is assumed to be a stochastic variable following a *categorical distribution*. The stochastic variable depends on i, but draws for different i are assumed to be independent. The unknown event probabilities are denoted by $p_{ghi} = \mathbb{P}(\widehat{s}_i = h \mid s_i = g)$ and stored in *confusion matrices* $P_i = (p_{ghi})_{gh}$ of dimension $K \times K$. The suggested estimator $\widehat{\boldsymbol{u}}$ is shown to have expectation $\mathbb{E}[\widehat{\boldsymbol{u}}] = \sum_{i \in I} P_i^T \widehat{\boldsymbol{a}}_{\boldsymbol{i}} y_i$, where $\widehat{\boldsymbol{a}}_{\boldsymbol{i}}$ is the i-th column of $\widehat{A}^T$. If P_i^T is invertible, we denote its inverse by Q_i. An unbiased estimator of $\boldsymbol{u}$ is then given by $\sum_{i \in I} Q_i \widehat{\boldsymbol{a}}_{\boldsymbol{i}} y_i$. We refer to this bias correction method as *the baseline method*.

What remains is to estimate the classification error rates p_{ghi} using a test set. In Section 4.3, we propose a novel estimation method that properly deals with small test sets (recall The Peculiar Example).

4.2.2 Related work

Below, we discuss related work on biased aggregates, bias in machine learning and Bayesian inference.

Biased Aggregates. To the best of our knowledge, the paper by Van Delden et al. (2016) is the first work *generally* describing the classification error model and studying classifier-based statistics.[1] We mention front runners from two fields using similar bias correction methods.

The first field is *epidemiology*, which studies the distribution of health and disease. There, it is well-known how a low base rate can lead to large bias even if sensitivity $(1 - p)$ and specificity $(1 - q)$ are high (Lash, Fox and Fink, 2009, cf. The Base Rate Example). As mostly binary classifiers are considered (sick or not), bias correction is straightforward (Lash et al., 2009, pp. 87-89). In addition, a standard Bayesian method of bias correction, predominantly using a uniform or Jeffreys prior without parameter constraints, is applied in epidemiology (Goldstein et al., 2016; Gustafson, 2004). We will generalise these Bayesian methods to the entire family of conjugate prior distributions and to the setting of multi-class

[1] After publication of this chapter, we discovered that the first work is by Bross (1954). Section 1.4 contains more details.

classification. Moreover, we will improve empirical performance by imposing well-chosen parameter constraints.

The second field is *land cover mapping*, which analyses land use based on large volumes of remote sensing data, for example using SVM (Löw et al., 2015). As can be expected, multi-class classification is not uncommon in this subject area (see Costa et al., 2018, for a case study with $K = 15$ classes). Since accurately estimating the total area of types of vegetation is highly relevant in monitoring ecological systems (Veran, Kleiner, Choquet, Collazo & Nichols, 2012), there have been many efforts in the field to make better use of accuracy data (Olofsson, Foody, Stehman & Woodcock, 2013). To the best of our knowledge, our bias correction method is a novel contribution to these efforts.

Bias in Machine Learning. In machine learning, the accuracy of classifier-based statistics is relatively understudied.[2] The literature on machine learning is mainly concerned with minimising loss for *individual* future predictions and therefore focuses on a different kind of bias. Much work deals with model selection bias and overfitting (Cawley & Talbot, 2010) and sample selection bias (Zadrozny, 2004). However, reducing these types of bias (by, e.g., using k-fold cross-validation) does not necessarily reduce the *statistical* bias of classifier-based statistics. This is a pitfall especially if the base rate is low, i.e., when dealing with *class-imbalanced data*. Of course, correctly dealing with class-imbalanced data is well-studied (Haixiang et al., 2017). However, as long as the classifier is not error-free, it will still result in biased aggregate predictions.

An alternative is an *aggregate loss function* that measures the bias on the aggregate level instead of on the individual level (Sodomka, Lahaie & Hillard, 2013). Other alternatives include averaging multiple (biased) estimators into a single, more accurate estimator (Taniguchi & Tresp, 1997). Such alternatives will reduce the bias of classifier-based statistics, but our proposed method will completely remove it for sufficiently large test sets.

Bayesian Inference. An extensive review on Bayesian inference for categorical data analysis is provided by Agresti and Hitchcock (2005). It specifically comments on the use of prior distributions and "the lack of consensus about what noninformative means" (Agresti & Hitchcock, 2005, p. 303). We will avert this

[2] We were alerted to the literature on quantification learning (see González et al., 2017) only after publication of this chapter. Nonetheless, a similar claim is made by them, see also Section 1.4.

discussion by analytically deriving the posterior distribution for the family of conjugate priors. We empirically evaluate two common prior choices: the uniform (flat) prior and the Jeffreys prior. In real-world applications, our bias correction method can be implemented for non-conjugate prior as well, by utilising Markov chain Monte Carlo (MCMC) methods.

4.3 Methods

In this section, we introduce our bias correction method in the general setting of multi-class classification problems. The section contains three parts. First, the likelihood and posterior (for conjugate priors) as well as the Jeffreys prior are shown. Second, we formulate novel constraints on the classification error rates, being our main scientific contribution. Third, we show how to obtain the proposed Bayesian estimator of the aggregate vector using these parameter constraints.

4.3.1 Bayesian parameter estimation

We begin by translating existing theorems from Bayesian statistics to our setting of multi-class classification. It mostly contains elementary probability manipulations. We make one simplifying assumption compared to Van Delden et al. (2016): the probabilities p_{ghi} do not depend on i. We write p_{gh} instead and only a single confusion matrix P remains. We refer to this assumption as that of the *homogeneity of the confusion matrices*. The reason for this simplifying assumption is that the notation, derivations and formulas are more pleasant to read. In practice, it might be more reasonable to assume that p_{ghi} depends on i, but only through y_i and possibly other features. Our proposed methodology can be applied separately to each group of objects having similar features, which can then be aggregated to obtain a single, final estimate.

Two parts now follow: (1) formulating the likelihood function and posterior distribution (for conjugate priors), and (2) deriving the Jeffreys prior.

Likelihood and Posterior. The parameters of the classification error model are the K^2 classification error rates (or *event probabilities*) $\{p_{gh} : g, h \in H\}$ and the K base rate parameters $\{\alpha_g : g \in H\}$. To estimate these parameters, consider a test set $J = \{i_1, \ldots, i_n\} \subset I$ of n randomly selected $i_j \in I$ for which we observe s_{i_j}. The corresponding data set $\mathcal{D}$ of n independent observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is given by $\boldsymbol{x}_j = (s_{i_j}, \widehat{s}_{i_j})$, for $i_j \in J$. We simply write $\boldsymbol{x}_j = (s_j, \widehat{s}_j)$ for $j = 1, \dots, n$. The following theorem shows the likelihood function and the posterior distribution for a suitable family of prior distributions, namely Dirichlet distributions. Recall that a Dirichlet distribution of order $k \geq 2$ has the standard $(k-1)$-simplex $\Delta_{k-1} \subset \mathbb{R}^k$ as support, where

$$\Delta_{k-1} = \left\{ \boldsymbol{x} \in \mathbb{R}^k : x_i \geq 0, \ \sum_i x_i = 1 \right\}. \tag{4.6}$$

The density of a Dirichlet distribution of order $k \geq 2$ with concentration parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ equals

$$f(\boldsymbol{x} \mid \boldsymbol{\beta}) = \frac{\Gamma(\beta_1 + \cdots + \beta_k)}{\Gamma(\beta_1) \cdots \Gamma(\beta_k)} \prod_{m=1}^{k} x_m^{\beta_m - 1}, \tag{4.7}$$

where $\Gamma(\cdot)$ is the gamma function. A Dirichlet distribution will be referred to as $\mathrm{Dir}(k, \boldsymbol{\beta})$.

Theorem 4.1. *The likelihood function of observing the data set $\mathcal{D} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, given the model parameters $\boldsymbol{p}$ and $\boldsymbol{\alpha}$, is given by*

$$p(\mathcal{D} \mid \boldsymbol{p}, \boldsymbol{\alpha}) = \prod_{g,h \in H} \left(p_{gh} \alpha_g\right)^{n_{gh}}, \tag{4.8}$$

where $n_{gh} = |\{j \in J : x_j = (g,h)\}|$. The family consisting of products of $K+1$ independent Dirichlet distributions of order K forms the collection of conjugate priors for the above likelihood function. Next, choose the prior $\prod_{k=1}^{K+1} D_k$ on $(\boldsymbol{p}, \boldsymbol{\alpha})$, where $D_g \sim \mathrm{Dir}(K, \boldsymbol{\beta}_g)$ for $g \in H$ and $D_{K+1} \sim \mathrm{Dir}(K, \boldsymbol{\gamma})$. Write $\boldsymbol{\beta}_g = (\beta_{gh})_{h \in H}$ for $g \in H$ and $\boldsymbol{\gamma} = (\gamma_g)_g$. The posterior density is then given by

$$p(\boldsymbol{p}, \boldsymbol{\alpha} \mid \mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \left(\prod_{g,h \in H} p_{gh}^{\beta_{gh} + n_{gh} - 1} \right) \left(\prod_{g \in H} \alpha_g^{\gamma_g + n_g - 1} \right), \tag{4.9}$$

where $n_g = \sum_{h \in H} n_{gh}$. Thus, the posterior distribution is the product of $K+1$ independent Dirichlet distributions, the first K being $\mathrm{Dir}(K, (\beta_{gh} + n_{gh})_{h \in H})$, for $g \in H$, and the last one being $\mathrm{Dir}(K, (\gamma_g + n_g)_{g \in H})$.

Proof. The probability that $s_i = g$ for a randomly selected $i \in I$ equals α_g. The probability of observing $\boldsymbol{x}_i = (g, h)$, given the classification error model, is equal to

$$\mathbb{P}(s_i = g, \widehat{s}_i = h) = \mathbb{P}(\widehat{s}_i = h \mid s_i = g)\,\mathbb{P}(s_i = g) = p_{gh}\alpha_g. \tag{4.10}$$

The pairs $\boldsymbol{x}_i$ again follow a categorical distribution, now mapping into the product space $\{1, \ldots, K\} \times \{1, \ldots, K\}$ having event probabilities $p_{gh}\alpha_g$. Equation (4.8) now follows (see Bishop, 2006, pp. 74-75).

Next, we observe that the likelihood function in expression (4.8) is equal to

$$p(\mathcal{D} \mid \boldsymbol{p}, \boldsymbol{\alpha}) = \left(\prod_{g,h \in H} p_{gh}^{n_{gh}} \right) \left(\prod_{g \in H} \alpha_g^{n_g} \right), \tag{4.11}$$

where $n_g = \sum_h n_{gh} = |\{j \in J : s_j = g\}|$. The likelihood function can be viewed as a product of $K + 1$ independent categorical distributions, as $\sum_h p_{gh} = 1$ for every $g \in H$. It is well-known that the family of Dirichlet distributions forms the collection of conjugate priors to the categorical (and multinomial) distribution. For a derivation, see Bishop (2006, pp. 76-78). This proves the claim regarding the family of conjugate priors for the likelihood function (4.8).

Finally, to derive the posterior distribution, we apply the result from Bishop (2006, pp. 76-78) to each of the parameter vectors $\boldsymbol{p}_g$ (with $g \in H$) and $\boldsymbol{\alpha}$ separately. The posterior distribution for a prior distribution $\prod_{k=1}^{K+1} D_k$ on $(\boldsymbol{p}, \boldsymbol{\alpha})$, where $D_g \sim \text{Dir}(K, \boldsymbol{\beta}_g)$ for $g \in H$ and $D_{K+1} \sim \text{Dir}(K, \boldsymbol{\gamma})$, is then seen to be equal to

$$p(\boldsymbol{p}, \boldsymbol{\alpha} \mid \mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \left(\prod_{g,h \in H} p_{gh}^{\beta_{gh} + n_{gh} - 1} \right) \left(\prod_{g \in H} \alpha_g^{\gamma_g + n_g - 1} \right).$$

This concludes the proof. □

The Jeffreys Prior. If the model is correctly specified, the effect of prior choices will diminish as n attains larger values. For small test set sizes n, the choice of a prior distribution does affect the posterior distribution. With the breakthrough of MCMC methods, there is no need to impose any restrictions (other than those resulting from numerical limitations) on the family of prior distributions. For now, we have considered conjugate prior distributions in order to analytically formulate the posterior distribution. Specifically, we will compare two common prior choices: (1) the uniform prior and (2) the Jeffreys prior (cf. Agresti & Hitchcock, 2005). The uniform prior corresponds to setting the components of the hyperparameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_g$, $g \in H$, equal to 1 in Theorem 4.1. The Jeffreys prior, defined as proportional to the square root of the determinant of the Fisher Information Matrix (FIM), corresponds (in the case of a single categorical stochastic variable) to setting all components of $\boldsymbol{\beta}$ equal to 1/2 (Agresti & Hitchcock, 2005, p. 303). However, due to the repeated occurrence of α_g in the likelihood function (4.8), we find a slightly different outcome which we did not find in recent text books.

Proposition 4.1. *The Jeffreys prior for the likelihood function* (4.8) *is given by a product of* $K+1$ *independent Dirichlet distributions of order* K *with hyperparameters* $\beta_{gh} = 1/2$ *for all* $g, h \in H$ *and* $\gamma_g = K/2$ *for all* $g \in H$.

Proof. The proof consists of computing the determinant of the Fisher information matrix (FIM). To compute the FIM, a linearly independent set of parameters specifying the model is required. Note that the parameter vectors $\{\boldsymbol{p}_g : g \in H\}$ and $\boldsymbol{\alpha}$ are elements of the K-simplex Δ_K, as $p_{gg} = 1 - \sum_{h \neq g} p_{gh}$ for each $g \in H$ and $\alpha_K = 1 - \sum_{g \neq K} \alpha_g$, where we identify H with the set $\{1, 2, \ldots, K\}$. It follows that the classification error model has $(K+1)(K-1) = K^2 - 1$ free parameters, which we will denote by $\widetilde{\boldsymbol{p}} = (\widetilde{\boldsymbol{p}}_g)_{g \in H}$, with $\widetilde{\boldsymbol{p}}_g = (p_{gh})_{h \neq g}$, and $\widetilde{\boldsymbol{\alpha}} = (\alpha_g)_{g \neq K}$. It is straightforward to show that the FIM is a block-diagonal matrix of the form

$$\boldsymbol{I}(\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{\alpha}}) = \begin{pmatrix} A_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & & \vdots \\ \vdots & & A_K & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & A_{K+1} \end{pmatrix}, \tag{4.12}$$

where $A_g = \alpha_g \left[p_{gg}^{-1} \cdot \mathbf{1} \cdot \mathbf{1}^T + \operatorname{diag}(\widetilde{\boldsymbol{p}}_g)^{-1} \right]$ for $g \in H$ and $A_{K+1} = \alpha_K^{-1} \cdot \mathbf{1} \cdot \mathbf{1}^T + \operatorname{diag}(\widetilde{\boldsymbol{\alpha}})^{-1}$, see Hogg, McKean and Craig (2018, p. 391). We will now use the fact that the determinant of an $m \times m$ matrix $C = \beta \boldsymbol{a}\boldsymbol{b}^T + \operatorname{diag}(\boldsymbol{d})$, with $\beta \in \mathbb{R}$ and $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{d} \in \mathbb{R}^m$ is given by

$$\det(C) = \left(1 + \beta \sum_{j=1}^{m} \frac{a_j b_j}{d_j}\right) \prod_{j=1}^{m} d_j. \tag{4.13}$$

For a proof, consult Graybill (1983, pp. 293-294). Applied to (4.12), we find $\det(A_g) = \alpha_g^{K-1} \prod_{h \in H} p_{gh}^{-1}$, for $g \in H$ and $\det(A_{K+1}) = \prod_{g \in H} \alpha_g^{-1}$. It follows that

$$\det(\boldsymbol{I}(\boldsymbol{p}, \boldsymbol{\alpha})) = \left(\prod_{g,h \in H} p_{gh}^{-1} \right) \left(\prod_{g \in H} \alpha_g^{K-2} \right). \tag{4.14}$$

Taking the square root concludes the proof. □

The Jeffreys prior density for the binary classification problem ($K = 2$) reduces to

$$p(\boldsymbol{p}, \boldsymbol{\alpha} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{\pi^2 \sqrt{p(1-p)q(1-q)}}, \tag{4.15}$$

where $p = p_{10}$ and $q = p_{01}$. Panel (A) of Fig. 4.1 shows the marginal Jeffreys prior density for the parameter $p = p_{10}$ (1 minus recall) for $K = 2$. Panels (B) – (D) show the marginal posterior density for $n = 50$, $n = 500$ and $n = 2{,}000$, using simulated data with base rate $\alpha_1 = 0.1$ and true classification error rates $p = 0.3$ and $q = 0.1$. The posterior density indeed converges to the true parameter value $p = 0.3$ (dashed line).

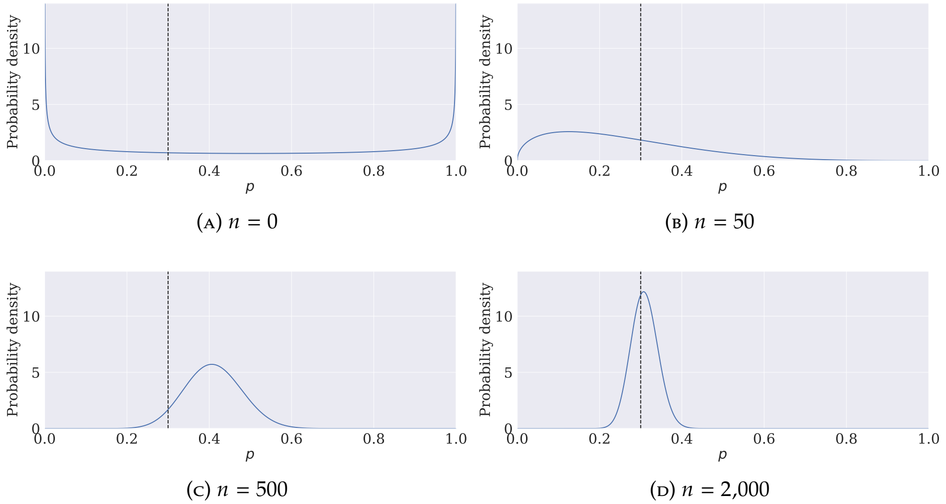
4.3.2 Imposing parameter constraints

Motivated by The Peculiar Example, we deduce useful constraints for the model parameter $\boldsymbol{p}$. The example showed that in (very) small test sets, the unbiased estimator $Q\widehat{\boldsymbol{v}}$ of the counts vector $\boldsymbol{v}$ might give impermissible estimates (negative counts). In particular, we estimated the true webshop count to be −75.

In our setting, there are two possible explanations for finding a negative estimate of a positive quantity: (I) the predicted base rate vector, being only a single realisation of a stochastic variable, lies far away from its mean in this particular case, or (II) we have estimated the confusion matrix not sufficiently accurately and therefore the bias correction is inaccurate. The following theorem demonstrates why, in many practical cases, (II) plays a larger role than (I).

Theorem 4.2. *The variance-covariance matrix of the base rate vector $\widehat{\boldsymbol{\alpha}}$, conditional on the true s_i, equals*

$$V(\widehat{\boldsymbol{\alpha}} \mid \{s_i, i \in I\}) = \frac{1}{N} \left(\operatorname{diag}(P^T \boldsymbol{\alpha}) - P^T \operatorname{diag}(\boldsymbol{\alpha}) P \right). \tag{4.16}$$



(A) $n = 0$

(B) $n = 50$

(C) $n = 500$

(D) $n = 2{,}000$

Fig. 4.1: The Jeffreys prior (without parameter constraints) and the resulting posterior densities for p, converging to the true parameter $p_0 = 0.3$ as n increases.

Proof. Recall the $N \times K$ class matrix $A = (a_{ih})$ given by $a_{ih} = I(s_i = h)$ for objects $i \in I$ and classes $h \in H$. The vector $\boldsymbol{a_i}$ denoted the i-th column of A^T. Note that the equality $\boldsymbol{a_i}\boldsymbol{a_i}^T = \operatorname{diag}(\boldsymbol{a_i})$ holds, as each $\boldsymbol{a_i}$ is a standard basis vector of $\mathbb{R}^K$. A similar equality holds for $\widehat{\boldsymbol{a}}_{\boldsymbol{i}}$. The variance-covariance matrix of $\widehat{\boldsymbol{a}}_{\boldsymbol{i}}$ (for readability, we leave out the conditionality on the right-hand side) is equal to

$$\begin{aligned} V(\widehat{\boldsymbol{a}}_{\boldsymbol{i}} \mid \{s_i, i \in I\}) &= \mathbb{E}(\widehat{\boldsymbol{a}}_{\boldsymbol{i}}\widehat{\boldsymbol{a}}_{\boldsymbol{i}}^T) - \mathbb{E}(\widehat{\boldsymbol{a}}_{\boldsymbol{i}})\,\mathbb{E}(\widehat{\boldsymbol{a}}_{\boldsymbol{i}})^T \\ &= \mathbb{E}(\operatorname{diag}(\widehat{\boldsymbol{a}}_{\boldsymbol{i}})) - P^T\boldsymbol{a_i}\boldsymbol{a_i}^T P \\ &= \operatorname{diag}(P^T\boldsymbol{a_i}) - P^T \operatorname{diag}(\boldsymbol{a_i})P. \end{aligned}$$

In the last equality, we used the fact that the operation $\operatorname{diag}(\cdot)$ commutes with taking the expectation. Recall that $\widehat{\boldsymbol{a}}_{\boldsymbol{i}}$ and $\widehat{\boldsymbol{a}}_{\boldsymbol{j}}$ $(i \neq j)$ are independent, conditional on $\{s_i, i \in I\}$. Thus, summing both sides of the above equation over $i \in I$ and dividing the results by N^2 concludes the proof. □

Theorem 4.2 shows that the variance of $\widehat{\boldsymbol{\alpha}}$ is proportional to the inverse of the *population size* N, while the variance of the parameter $\boldsymbol{p}$ is proportional to the

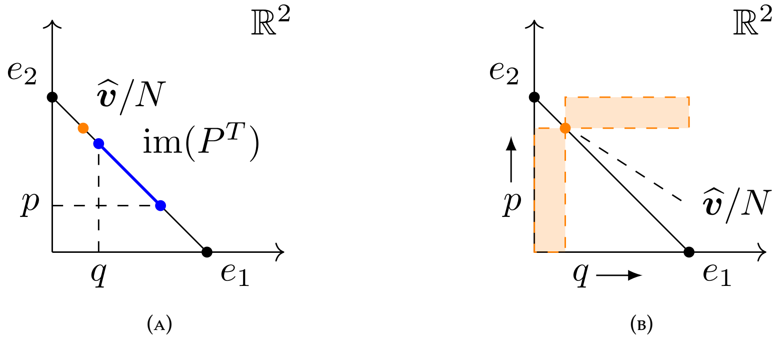


FIG. 4.2: The geometric relations between the confusion matrix P, the estimated counts vector $\widehat{\boldsymbol{v}}$ and the size N of the unseen data are shown in panel (A), for $K = 2$. Panel (B) shows the constraints on the model parameters p and q that we impose.

inverse *test set size* n. In practice, we often find $N \gg n$, hence (II) plays a larger role than (I).

Therefore, we wish to impose constraints on the model parameter $\boldsymbol{p}$ such that $Q\widehat{\boldsymbol{v}} \geq \boldsymbol{0}$ with probability 1, conditional on $\widehat{\boldsymbol{v}}$. Given $\widehat{\boldsymbol{v}}$, we write

$$\Theta_K = \Theta_K(\widehat{\boldsymbol{v}}) := \left\{\boldsymbol{p} \in (\Delta_{K-1})^K : Q(\boldsymbol{p})\widehat{\boldsymbol{v}} \geq \boldsymbol{0}\right\}, \tag{4.17}$$

where Δ_{K-1} is the standard $(K-1)$-simplex in $\mathbb{R}^K$. In other words, to determine whether a point $\boldsymbol{p}$ is in Θ_K, one has to check whether or not $\widehat{\boldsymbol{v}}$ is contained in the convex hull of the rows of the confusion matrix $P(\boldsymbol{p})$ corresponding to the point $\boldsymbol{p}$.

The Binary Case. In the binary case of $K = 2$, the proposed parameter constraints take on the elegant form

$$\Theta_2(\widehat{\boldsymbol{v}}) \cong ([0, \widehat{v}_2] \times [0, \widehat{v}_1]) \cup ([\widehat{v}_2, 1] \times [\widehat{v}_1, 1]). \tag{4.18}$$

To prove Equation (4.18) algebraically, solve the two linear equations $Q(\boldsymbol{p})\widehat{\boldsymbol{v}} \geq \boldsymbol{0}$ for $p = p_{10}$ and $q = p_{01}$.[3] Instead of doing so, we prefer a more insightful geometrical

[3] After publication of this chapter, we discovered the work by Molinari (2008). She derives Equation (4.18) as identification region as well. Moreover, she proves for arbitrary dimensions that the identification regions are star convex and provides a figure similar to Fig. 4.2, panel (B).

proof, see Fig. 4.2, in which e_1 and e_2 correspond to the standard basis vectors of $\mathbb{R}^2$. Note that the predicted base rate vector $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{v}}/N$ (orange dot) lies on the 1-simplex $\Delta_1 \subset \mathbb{R}^2$. The blue line segment in Fig. 4.2, panel (A), shows the image of Δ_1 under P^T and the corresponding locations of p and q on the axes. It shows precisely what we found in The Peculiar Example: the predicted counts vector does not lie in the image of Δ_1 under the *estimated* transposed confusion matrix P^T. Now, Theorem 4.2 shows that it is unlikely that the predicted base rate vector $\widehat{\boldsymbol{\alpha}}$ is not contained in the image of the *true* transposed confusion matrix. In Fig. 4.2, panel (A), we thus impose that we need a blue endpoint of $\mathrm{im}(P^T)$ on each side of the orange dot $\widehat{\boldsymbol{v}}/N$. It follows that the orange area in Fig. 4.2, panel (B), contains the permitted (q, p)-pairs and thus corresponds to $\Theta_2(\widehat{\boldsymbol{v}})$. This concludes the geometrical proof of Equation (4.18).

4.3.3 Bayesian aggregates

Van Delden et al. (2016) showed that $\mathbb{E}[\widehat{\boldsymbol{u}}] = P^T\boldsymbol{u}$. Hence, an unbiased estimator of $\boldsymbol{u}$ would be $Q\widehat{\boldsymbol{u}}$ with $Q = (P^T)^{-1}$. Recall that we refer to this method as the baseline method. The baseline method does not yet take the uncertainty in estimating the confusion matrix P into account. In The Peculiar Example, we have seen that this might lead to impermissible outcomes for small test sets. We propose the following three-step approach, assuming that a classification algorithm has already been trained and used to predict the classifications on the entire unseen data set. In step 1, choose a prior from the Dirichlet family, such as the Jeffreys prior in Proposition 4.1, and find the posterior distribution of $\boldsymbol{p}$ using Theorem 4.1 (integrating $\boldsymbol{\alpha}$ out of the equation). Alternatively, MCMC methods can be used to obtain a numerical approximation of the posterior distribution, also for non-conjugate (but proper) priors. In step 2, take draws from the posterior distribution and only accept the draws that lie within $\Theta_K(\widehat{\boldsymbol{v}})$. Choose a positive integer R (resolution parameter) and stop step 2 after R draws have been obtained. In step 3, compute the matrix Q and the product $Q\widehat{\boldsymbol{u}}$, for each of the R draws. These three steps give a numerical approximation (of resolution R) to the posterior distribution of the estimator of the aggregate vector $\boldsymbol{u}$, conditional on $\widehat{\boldsymbol{u}}$.

We conclude by noting that the time complexity of the proposed Bayesian method in n and N is equal to that of the baseline method if conjugate priors are considered, as it allows direct sampling from the posterior distribution. In fact, if wall-clock time is considered, the entire estimation can be easily implemented

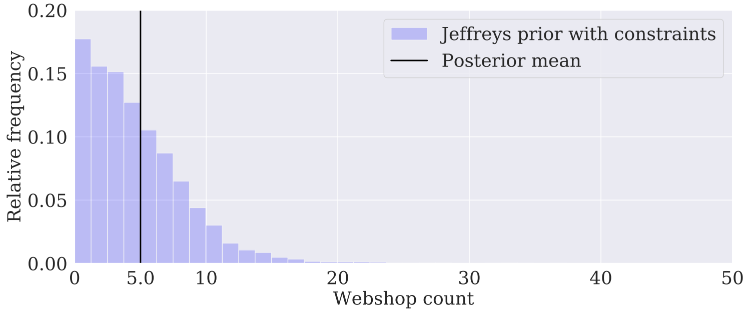


FIG. 4.3: The results of our method (with Jeffreys prior and parameter constraints) for The Peculiar Example.

to finish in tens of milliseconds on any regular machine. Using MCMC methods, the time complexity depends on the prior choice and might increase with n, but not with N. The wall-clock time for our bias correction method using MCMC methods will increase to several minutes.

4.4 Empirical evaluation

We begin by briefly revisiting The Peculiar Example from Section 4.1. Then, we introduce real-world data on company tax returns and use it to compare existing methods to our bias correction method.

4.4.1 A solution for The Peculiar Example

Recall The Peculiar Example from Section 4.1. The baseline method resulted in an estimate of −75 webshops. The result of our bias correction method, with Jeffreys prior and parameter constraints, is shown in Fig. 4.3. The posterior mean of the number of items in the webshop class is now equal to 5.0. The distribution is skewed (to the right) and the estimator still has a large variance, indicating that more labelled data are required to obtain a more accurate estimate. However, our bias correction method *is* able to capture the little information available in the test set of size $n = 10$, resulting in a permissible estimate.

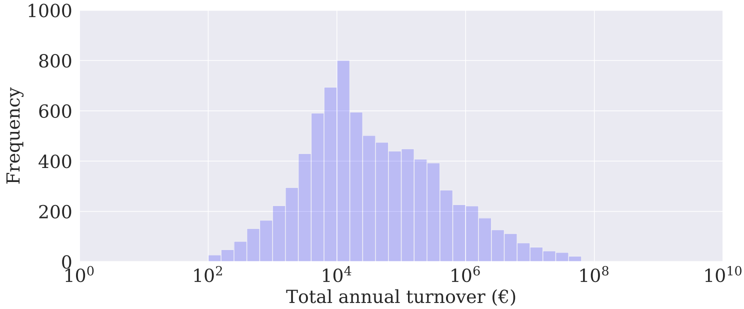


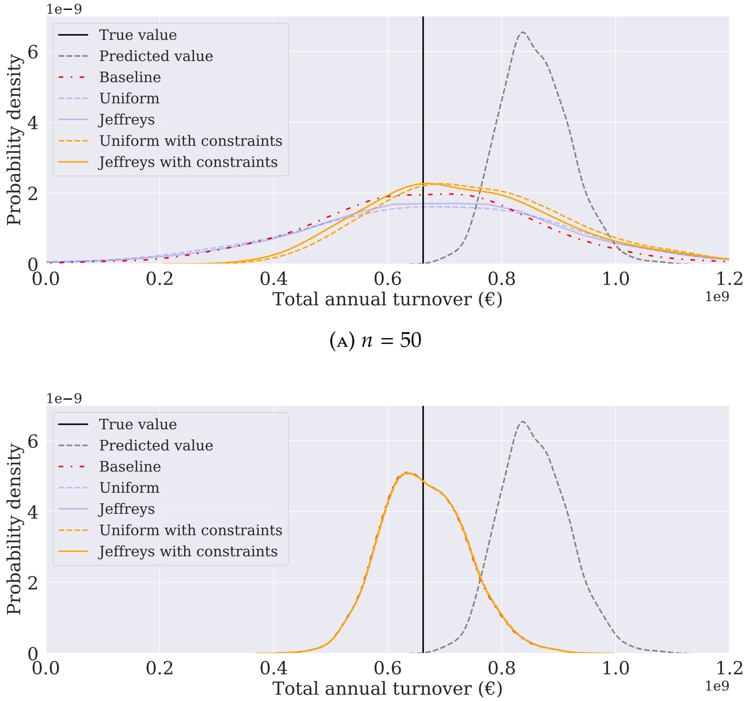
FIG. 4.4: The distribution of the turnover variable in a data set of filed tax returns. Figure adapted from Meertens et al. (2018), which was later published as Meertens et al. (2020) and corresponds to Chapter 5 of this thesis, see Fig. 5.4, panel (C).

4.4.2 Data on company turnover

To show the strength of our bias correction method in a general setting, we will now empirically evaluate our bias correction method using a real-world data set of company tax returns. The goal is to estimate the total annual turnover of webshops. The binary classification (webshop or not) is not available in the data, but the annual turnover y_i for each company i is. This application is exhaustive, because the numerical variable y (company turnover) is not constant (i.e., y_i depends on i).

The data set contains tax returns filed in the Netherlands in 2016 indexed by a set I of $N = 18{,}939$ companies established outside the Netherlands, but within the European Union (Meertens et al., 2018). The total turnover reported by these companies for 2016 equals EUR 12.2 billion. Fig. 4.4 shows how companies' total annual turnover is distributed in the data set.

Meertens et al. (2018) trained a classification algorithm (on a separate data set) to predict whether a company was active as a webshop (class 1) or not (class 0). Without discussing the details of training it, we run the classification algorithm on the full data set of size N, obtaining the predicted classifications $\widehat{s}_i \in \{1, 0\}$.



(A) $n = 50$

(B) $n = 2{,}000$

FIG. 4.5: The distribution of the mean of the posterior distribution of total turnover for different priors and different test set sizes.

4.4.3 Bias correction and effect of prior beliefs

We compare the accuracy of our bias correction method to that of the baseline method and that of Bayesian methods without parameter constraints, when applied to estimate the total turnover of webshops in the given data set. To compute the accuracy of the estimators, we have to know the true classification s_i (webshop or not) for each company. As the true classifications are unknown, the *only* way to examine the accuracy is by means of simulation. For the purpose of this simulation, we take the predicted classifications from Meertens et al. (2018) as the true classifications s_i. The base rate of the webshop class then equals $\alpha_1 = 0.075$.

Next, we take the binary classification error model with the true classification error rates $p = 0.05$ and $q = 0.05$ as the data-generating process. We make this

Table 4.1: Comparison of the bias, variance and MSE of each of the bias correction methods (including no correction) when estimating webshop turnover.

	n = 50			n = 2,000		
	Bias	Var	MSE	Bias	Var	MSE
Correction method	$\times 10^{6}$	$\times 10^{15}$	$\times 10^{15}$	$\times 10^{6}$	$\times 10^{15}$	$\times 10^{15}$
None	195,3	3,7	41,8	195,3	3,7	41,8
Baseline	-1,3	46,3	46,3	1,3	5,4	5,4
Uniform	10,8	72,7	72,8	3,1	5,5	5,5
Jeffreys	22,0	81,7	82,2	2,3	5,5	5,5
Uniform with constraints	105,1	35,8	46,9	3,1	5,5	5,5
Jeffreys with constraints	88,5	39,0	46,9	2,3	5,5	5,5

choice of p and q, because it leads to a sufficiently large bias of classifier-based statistics (as $q/(p+q) = 0.5 \gg 0.075 = \alpha_1$) making Fig. 4.5 more pleasant to read. We emphasise that other choices of p and q will lead to similar conclusions.

The results of our Monte Carlo simulation are shown in Fig. 4.5, including the true value of webshop turnover u_1 (black, solid) and the distribution of the predicted value $\widehat{u}_1$ without performing any corrections (black, dashed). Besides our bias correction method (orange, dashed and solid), the figure shows the baseline method (red, dashed) and existing Bayesian bias correction methods without parameter constraints (blue, dashed and solid). The distributions are obtained by drawing 1,000 bootstrap replications from the classification error model and following the approach proposed in Section 4.3.3 for each bootstrap replication. Panel (A) shows the results for a test set of size n = 50 and panel (B) shows the results for n = 2,000.

To facilitate a more rigorous comparison of the methods, the same results are summarised in Table 4.1. We make four observations from the results. First, for n = 50, our bias correction method achieves a considerable reduction of the MSE compared to existing Bayesian methods without parameter constraints. The effect of imposing parameter constraints diminishes for n = 2,000. Second, for n = 2,000, our bias correction method performs equally well (in terms of the MSE) as the baseline method and existing Bayesian methods. Third, also for n = 2,000, all bias correction methods are a huge improvement (in terms of the MSE) compared to performing no bias correction; the bias substantially decreases without increasing the variance too much. Fourth, our bias correction method decreases the variance compared to the baseline method, essentially by cutting off a large part of the support (see Fig. 4.5, panel (A)). Even though this slightly increases the MSE, it guarantees that impermissible estimates are never found, as we have illustrated in Section 4.4.1.

Finally, we note that a data set of size $N = 18{,}939$ is rather small for data mining applications nowadays. However, The Base Rate Example and Equation (4.2) showed that the relative bias does not depend on N. Therefore, the experimental results that we have found for the data set of company tax returns will generalise to (much) larger data sets in other applications.

4.5 Chapter conclusions

In this paper, we have studied the statistical bias of classifier-based statistics in the general setting of multi-class classification. We proposed a Bayesian bias correction method, being the first to derive and impose constraints on the classification error rates.

For small test sets, imposing these parameter constraints dismisses impermissible estimates, leading to a similar MSE as the baseline method and a reduced MSE compared to existing Bayesian methods. For larger test sets, all bias correction methods yield similar results, substantially reducing the MSE of classifier-based statistics. The improvement of our method compared to existing methods is particularly compelling if the number of labelled data points is small. We argue this to be relevant in many data mining applications, including sentiment analysis and land cover mapping.

As future work, we aim to reduce the MSE even further by studying the bias-variance trade-off in classifier-based statistics in more detail, and by relaxing the assumption of homogeneity of confusion matrices.

CHAPTER 5

CROSS-BORDER INTERNET PURCHASES

5.1 Introduction

The accurate estimation of cross-border on-line consumption has recently become more important for two reasons. First, consumption through on-line channels of both goods and services is increasing within the European Union (EU), especially across borders. More consumers have access to the Internet, shipping costs are decreasing and payment services are converging across countries (Cardona & Duch-Brown, 2016; Marcus & Petropoulos, 2016; Martikainen, Schmiedel & Takalo, 2015). Accurate estimates of cross-border on-line consumption are of increasing importance for adequately reporting on national accounts by national statistical institutes. Second, cross-border on-line trade is nowadays a highly relevant item on the EU Digital Single Market policy agenda (European Commission, COM/2010/0245). Therefore, getting a grip on cross-border on-line consumption through reliable estimates is essential to for quantifying the effect of new policies. The need for accurate estimates of indicators of the digital economy within the EU is also emphasised by the European Commission (2015).

5.1.1 Existing survey-based approaches

Existing approaches for estimating consumption are based on either consumer surveys or business surveys. One EU-wide consumer survey on cross-border online consumption is conducted by Ecommerce Europe (`https://www.ecommerce-europe.eu`). In the Netherlands, this survey is conducted by market research institute the Gesellschaft für Konsumforschung (GfK) (`https://www.gfk.com`). It is commissioned by Thuiswinkel.org[1] on behalf of Ecommerce Europe. The estimates of total cross-border on-line consumption are based on asking consumers how much they spent at foreign webshops over a fixed time period in the past.

We argue that such an approach based on consumer surveys will lead to an underestimation of cross-border on-line consumption. We start our argumentation with the observation as made by Gomez-Herrera, Martens and Turlea (2014). They showed that one of the main impediments of on-line consumption by consumers within the EU is foreign language, rather than security reasons, shipping costs, geographical distance or available payment services. Consequently, webshops that are selling goods or services in multiple countries typically operate in a country by using a website in the regional language (Schu & Morschett, 2017).

[1] The national e-commerce association in the Netherlands, see `https://www.thuiswinkel.org`.

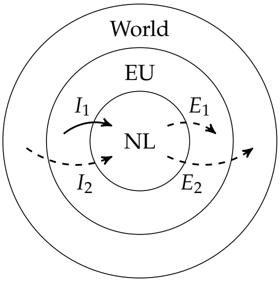


Fig. 5.1: The four types of cross-border flows of goods crossing an EU member state, e.g., the Netherlands (NL). The paper focuses on flow I_1 (solid arrow), the import of goods from other EU member states.

Therefore, a consumer cannot distinguish between domestic and foreign webshops, as both will be presented in their regional language.

Hence, we may conclude that a consumer survey approach leads to a downward bias in measuring cross-border on-line consumption. We shall refer to the downward bias of consumer survey approaches as *language bias*. To the best of our knowledge, language bias has only been pointed out before by Minges (2016), who was mainly concerned by the implications for official statistics. Here, we stress the scientific implication: current studies on cross-border on-line consumption and trade within the EU, mostly based on consumer survey data, might draw biased conclusions. We suggest that future studies on cross-border on-line consumption and trade should not be based on data obtained (solely) from consumer surveys. To support that suggestion, the goal of this paper is to construct a new and reliable methodology to obtain more accurate estimates of cross-border on-line consumption within the EU. Here, we initially focus on the consumption of goods (see Fig. 5.1).

The first reaction to addressing language bias is to use business surveys instead of consumer surveys (Minges, 2016). However, we believe that this would be unsatisfactory, for two reasons. First, measuring cross-border on-line consumption of consumers in a single country by business surveys requires large companies within the EU to report their sales to consumers per EU member state. This places a large administrative burden on companies. Second, the approach poses significant challenges to any correction for sampling probabilities and biases if, for example, the population of the existing EU-wide information and communication technologies (ICT) survey for enterprises (https://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm) would be used. The referenced population is the result of sampling and stratification with respect to (a) economic activity and (b) either relative turnover or number of employees. The stratified sampling probabilities with respect to size in one country must be transformed into that of the total on-line sales in another country. Given large differences between countries in this regard, it seems infeasible to arrive at accurate estimates of cross-border on-line consumption for each EU member state by using business surveys. In summary, both existing official statistical approaches are inadequate in estimating cross-border on-line consumption.

5.1.2 Our novel approach

The shortcomings of existing approaches that were discussed above are mostly due to the use of inadequate sources of data. Therefore, on the basis of our findings so far, we believe that data used for estimating cross-border on-line consumption should at least meet the following three requirements. First, the data must be based on supply-side information, preventing the aforementioned language bias. Second, the data must be collected for accurately measuring the sales of companies across borders. A reliable administrative or other integral data source would be preferable, since such data prevent having to deal with sampling issues. Third, the data must be available to national statistical institutes across the EU, allowing harmonised estimation methods across member states.

Motivated by these three requirements, we propose using tax returns filed by foreign companies (subsequently referred to as *tax data*). The EU system of value added tax (VAT) states that any company that is both established in the EU and involved with cross-border intra-community supplies to consumers must pay VAT in the country of destination and file a tax return (European Commission, Council Directive 2006/112/EC). The threshold value on total turnover from sales to consumers, above which filing a tax return is mandatory, is either €35,000 or €100,000, depending on the country of destination. For foreign companies selling to consumers in the Netherlands, the threshold value on total sales in the Netherlands equals €100,000. The Dutch Tax and Customs Administration collects such tax returns, which are then made available to Statistics Netherlands. Other EU member states will have similar, but not the same, data collection procedures, which we shall not discuss. We emphasise that using tax data restricts

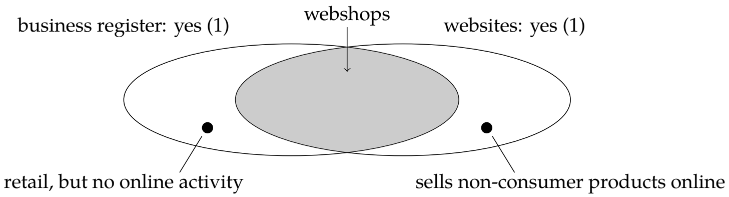

business register: yes (1)
webshops
websites: yes (1)
retail, but no online activity
sells non-consumer products online

FIG. 5.2: Combining the predictions by the business register with those by websites.

us to measuring cross-border *on-line consumption of goods* (henceforth referred to as *cross-border Internet purchases*).

The main challenge in using tax data is to identify webshops. For this identification, we propose an approach consisting of three steps. In the first step, the aim is to select the companies that are economically active in retail trade, according to the *nomenclature statistique des activités économiques dans la Communauté Européenne* (NACE), revision 2. (NACE, Rev. 2, is the statistical classification of economic activities in the EU, see also `http://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF`.) This is achieved by probabilistic record linkage (at firm level) of the tax data with a business register of retail companies that are active in the EU. In the second step, we use website data (obtained by web scraping) to confirm or complement the result from the first step. We apply machine learning in both of the first two steps to maximise the accuracy of the predictions. The results from the first two steps are combined by taking the intersection of the results (see Fig. 5.2). In the third step, we implement recently developed bias correction techniques (see Fig. 5.3) that have hitherto been overlooked by the machine learning community.[2] Below, we discuss our methodological contributions to the extant literature in each of the three steps, leading to our main contribution.

5.1.3 Related work and methodological contributions

Here, our methodological contributions to the scientific literature on (1) probabilistic record linkage, (2) web scraping and (3) machine learning are discussed.

[2] We were alerted to the literature on quantification learning only after publication of this chapter. Nonetheless, González et al. (2017) do mention that “quantification learning is still relatively unknown even to several machine-learning experts”.

The main scientific contribution is the contribution of the three methods and their incorporation in official statistics, which we point out subsequently.

Probabilistic Record Linkage. Firm-level record linkage in the absence of unique identifiers occurs often in economic research. One of the first well-known examples is the National Bureau of Economic Research's Patents Data Project (Hall, Jaffe & Trajtenberg, 2001). There, the matching was done mostly manually, which was "one of the most difficult and time-consuming tasks of the entire data construction project". Since that study, many automated alternatives, using approximate string matching algorithms, have been suggested. Recent examples include Bena, Ferreira, Matos and Pires (2017) and Tarasconi and Menon (2017). The general term for these approximate methods is *probabilistic record linkage*, first proposed in the seminal work by Fellegi and Sunter (1969). Now, two issues arise in applications of approximate string matching algorithms: (a) how to choose a similarity measure and (b) how to choose an optimal similarity threshold.

Issue (b) is an optimisation problem, that can be solved by using machine learning. Balsmeier et al. (2018) proposed to use k-means clustering for this. We shall improve on this work by comparing a wider range of machine learning algorithms and selecting the one that best fits the data. To overcome issue (a) we suggest combining the results of multiple similarity measures. These results can be used as features in the machine learning algorithm that optimises the similarity threshold to choose.

An interesting alternative is to use data from the Internet, as suggested by Autor, Dorn, Hanson, Pisano and Shu (2017). In brief, their approach entails finding the Uniform Resource Locator (URL) of a company's website by entering the company name into the Internet search engine Bing.com and then using the URL as a unique identifier in the matching process. The advantage of such an approach is that it does not use string matching algorithms at all but instead assumes that the variations in spelling of company names are stored in the database of the Internet search engine. Therefore, it solves the two issues of using string matching algorithms at once. However, there are at least two disadvantages of the approach. First, entering a legal company name in an Internet search engine might not always imply that the company's website is included in the top results, in particular for smaller companies. A second disadvantage is that the results are more difficult to reproduce, as the Internet is a dynamic source of data. We therefore propose a combination of Internet search (in our second step) and string

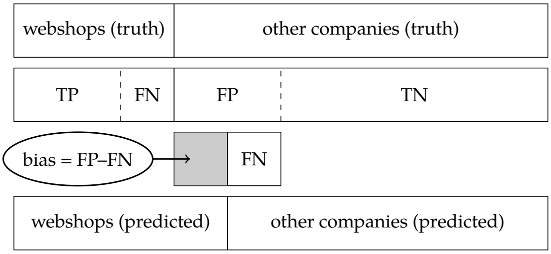


FIG. 5.3: Initial distribution (top row) of true labels (webshop or other company) and the final distribution (second and bottom row) of predicted labels. The bias in the predicted number of webshops (grey square) results from a difference in the number of false positive, FP, and false negative, FN, predictions, showing that the bias is 0 if and only if precision equals recall.

matching techniques (in the first step), so that we can benefit from the advantages of both approaches.

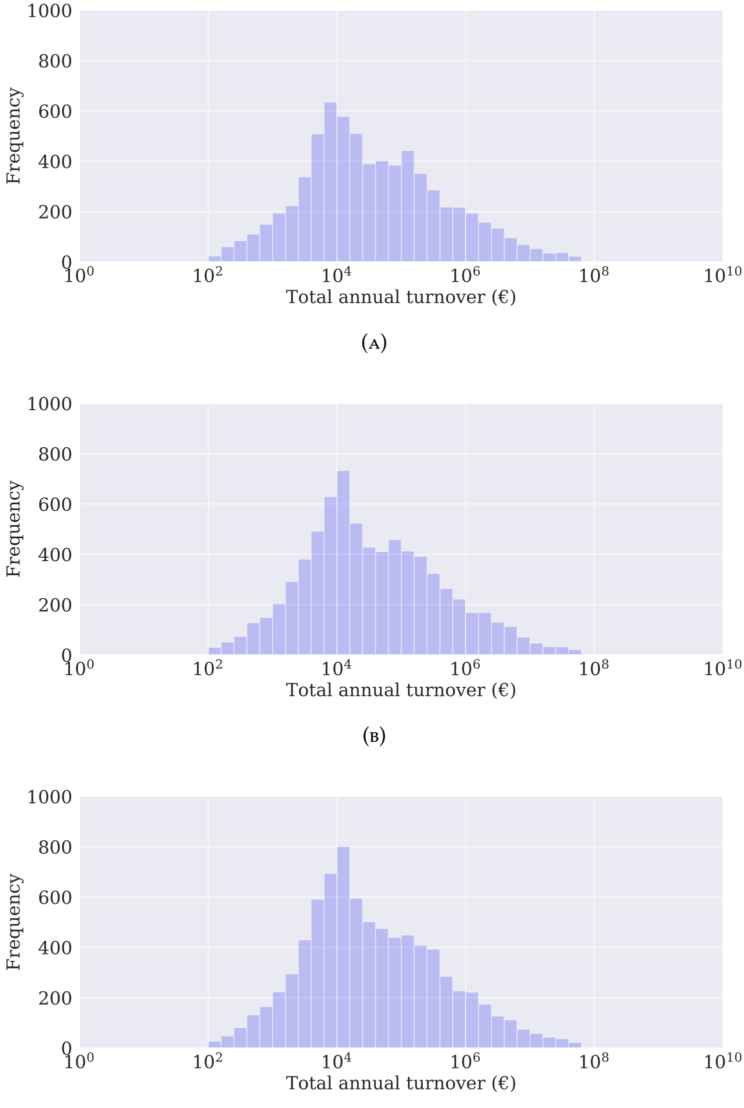
Web Scraping. In contrast with our fully data-driven work on firm-level record linkage, our work on web-based e-commerce detection uses manually selected features based on expert knowledge, because it is easier to understand and implement. Moreover, we show that the accuracy of our approach is high, ruling out the need to implement highly advanced, data-driven web-based e-commerce detection algorithms that are currently cutting edge (Blazquez, Domenech, Gil & Pont, 2018). We do use a (pretrained) machine learning model to find the website of a company based on the legal company name. In this respect, we improve on Autor et al. (2017). In addition, similarly to the first step, we compare the goodness of fit of a wide range of machine learning algorithms that use (knowledge-based) features obtained by web scraping from company websites to predict whether a company is a webshop or not.

Machine Learning. The first two steps provide an accurate (binary) prediction of whether the company is a webshop or not for each company in the set of tax returns (see Fig. 5.2). What remains is to aggregate the sales of goods of the identified webshops to obtain an estimate of cross-border Internet purchases by Dutch consumers within the EU. However, this estimate will be biased in general, as Fig. 5.3 illustrates. The fact that classifier-based aggregates are biased is relatively understudied in machine learning. To put it more strongly, we believe that we are the first to voice this observation.[3] A related (and mathematically equivalent) problem has been studied before. For example, in epidemiology (Lash et al., 2009) and land cover mapping (Löw et al., 2015), the effect of classification errors on aggregate estimates has been studied extensively for over at least three decades. To the best of our knowledge, the only work that *generally* discusses this effect is Van Delden et al. (2016), admittedly in the field of official statistics (and not in the field of machine learning). In all fields, the same equation for the bias of classifier-based aggregates has been derived. Our contribution to machine learning is that we show that the fundamental work by Van Delden et al. (2016) can be applied to automated classification algorithms in machine learning as well, leading to far more accurate estimates in general.

Official Statistics. The main contribution of our paper is to propose a novel methodology to estimate cross-border Internet purchases within the EU, exploiting data and methods hitherto not used for this. We demonstrate that there is convincing evidence from the Netherlands to show that our methodology results in more accurate estimates than approaches based on consumer surveys. The implementation of our methodology in the entire EU could lead to harmonised and accurate estimates of cross-border Internet purchases, ultimately providing policy makers with more reliable information regarding the EU Digital Single Market policy agenda.

The remainder of this paper is organised as follows. In Section 5.2, we describe the data that were used. We also describe how we obtained the data sets that were used to train the machine learning algorithms. In Section 5.3, the data-driven methods to identify foreign webshops are discussed. In Section 5.4, we present the results of applying the approach to the Netherlands and we compare them with results from an existing consumer-based approach, demonstrating the severity of the language bias. Section 5.5 concludes by discussing implementations for other EU member states and possible future research directions.

[3] We were alerted to the literature on quantification learning only after publication of this chapter. The observation was first voiced by Forman (2005).



(A)

(B)

(C)

Fig. 5.4: Distributions of annual turnover of foreign companies that filed a tax return in the Netherlands for the years (A) 2014, (B) 2015 and (C) 2016. The horizontal axis has a logarithmic scale. Companies that reported negative or zero turnover are not shown. Because of privacy legislation, bins containing fewer than 20 companies have been removed.

5.2 Data

In this section we first describe the supply-side data sets (tax data, websites, and the business register) that we used. Then, we show how the training and validation data set were obtained. Finally, we discuss the test data set.

5.2.1 Supply-side data

Below we discuss three types of supply-side data, i.e., tax data, data from websites and data from the business register.

Tax data. The data that were used to measure cross-border Internet purchases are tax returns filed in the Netherlands by foreign companies that are established in the EU. These tax data contain legal company names and the annual turnover from sales in the Netherlands of goods taxed at low or high tariff (i.e., sales to consumers). The data are extracted from tax returns filed for 2014, 2015 and 2016. The data set contains 197,424 filed tax returns from 22,440 unique companies. These tax data from the Netherlands are not openly available, because of strict privacy legislation. Under severe restrictions (among others, anonymising) and obeying serious impositions (such as suppressing extreme values), we are permitted to present aggregated figures on the data. When relevant, we reveal the criterion for which we suppressed information. In Fig. 5.4 we show the distribution of the annual turnover for each of the years 2014, 2015 and 2016. Furthermore, Table 5.1 displays summary statistics of the tax data from the Netherlands. As the global (cross-border) e-commerce market is rather complicated, we start by making three observations to clarify which flows can and cannot be measured by using the data set of tax returns.

First, smaller sellers might remain unobserved because of the threshold on annual cross-border on-line sales to Dutch consumers of €100,000. This is particularly problematic as many such small sellers exist: a fact referred to as the long tail of electronic commerce (Bailey et al., 2008; Oestreicher-Singer & Sundararajan, 2012). Many such small sellers use marketplaces (e.g., Amazon) as intermediaries. Now, the distance sales of such marketplaces typically exceed the annual turnover threshold (of €100,000). Therefore, they must file tax returns and they show up in the data. Fig. 5.5 shows in more detail which sales can and cannot be observed by using tax data as primary source of data to estimate cross-border Internet purchases.

Table 5.1: Summary statistics (mean, median and 10th and 90th percentile of annual turnover) of the tax returns filed in the Netherlands by foreign companies in 2014, 2015 or 2016.[†]

Year	$\lvert I \rvert$	$\lvert I_0 \rvert$	$\lvert I_{<0} \rvert$	$\lvert I_{>0} \rvert$	*Mean*	*Median*	*10th percentile*	*90th percentile*
2014	16,023	8,969	86	6,968	1,904,860	25,987	1,755	1,365,217
2015	17,313	9,771	80	7,462	1,603,908	25,166	1,783	1,218,926
2016	18,939	10,626	104	8,209	1,488,351	24,790	1,879	1,143,156

[†] The set of companies filing a tax return in a certain year in denoted by I. The subscript denotes whether the annual turnover in the given period is equal to 0, negative or positive. The number of elements in a set X is denoted by $|X|$.

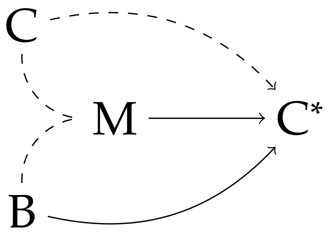


Fig. 5.5: Possible sales flows if a consumer buys goods on line, showing which flows can (solid) and cannot (dashed) be observed by using tax data (C, consumer (seller); B, business (seller); M, marketplace (seller); C*, consumer (buyer)). When the sales are facilitated by an on-line marketplace (e.g., Amazon), the distinction between sales by businesses and sales by consumers cannot be made. Therefore, the estimate of cross-border Internet purchases based on tax data might contain transactions from consumers to consumers.

Second, one might wonder to what extent Internet purchases at foreign companies that also have brick-and-mortar stores in the Netherlands show up in the data. Many such multinational companies exist, but most of them have organised their Internet sales from the Netherlands, as a result of which the monetary transaction from consumer to company does not cross borders. In some cases, the Internet sales are organised from outside the Netherlands. The consumer pays a foreign business entity and therefore the sales show up in the data, provided that the restrictions of threshold (€100,000 annual turnover) and location (established within the EU) are met. Hence, the trade flows that we observe coincide with the definition of import in the national accounts.

Third, it should be mentioned that our approach can only measure cross-border Internet purchases at companies established within the EU (see Fig. 5.1). Therefore, purchases at webshops located in, for example, China are not included in our estimates. This limitation is noteworthy, as the global e-commerce exports from China are growing vastly nowadays (S. Ma, Chai & Zhang, 2018).

Websites. The website of a company should be a clear indication of whether the company is a webshop or not. We shall use machine learning to distinguish websites of webshops from other websites. The hyper-text mark-up language (HTML) code of the home pages of the websites of companies are the data that we use as input. To obtain these data, we first select the legal company names of the 22,400 foreign companies that filed at least one tax return in the Netherlands for 2014, 2015 or 2016. Then, we use *URL retrieval* (Ten Bosch & Windmeijer, 2018) to find the home page of the websites belonging to these companies. Finally, we download the HTML code of the home pages from the Internet. The data were downloaded from the Internet on April 19th, 2017.

Business register. We used ORBIS as the business register, which is a global corporate database maintained by Bureau van Dijk (`http://bvdinfo.com/orbis`) and contains detailed corporate information on over 200 million private companies world wide. The database has been claimed to "suffer from some structural biases" (Ribeiro, Menghinello & De Backer, 2010). However, regarding European companies with an annual turnover of more than €100,000, the data set is practically complete (Garcia-Bernardo & Takes, 2018). Data from business registers regarding smaller foreign companies are not needed in our analysis, as these companies do not have to file tax returns in the Netherlands. The ORBIS database is used, because it contains the principal and secondary NACE (Rev. 2) codes for companies that are established in the EU. The NACE code can be used to select all active (and inactive) companies established in the EU and that are principally or secondarily economically active in retail trade. The result is a data set of 6,996,468 companies, from which companies established in the Netherlands have been excluded. This data set, including each company's country of establishment, was extracted from ORBIS on June 24th, 2017.

For our purposes, any business register containing the company names and country of establishment of every retail company in the EU would suffice, as long as the retail companies (according to NACE Rev. 2) can be identified as such. Therefore, we shall henceforth refer to the ORBIS data as the *business register*.

Table 5.2: Number of companies per industry class (using the Dutch statistical classification of industries from 1974) included in the training data set.[†]

Industry (1974)	*Threshold*	*Count*
Retail Trade	€ 1 million	100
Wholesale Trade	€20 million	30
Other	€50 million	50
Total	–	180

[†] The threshold value of the annual turnover is used as a selection criterion for a company to be included in the training data set.

5.2.2 Training and validation data set

In order to train classification algorithms, a labelled data set is required. Since no such data set existed, we manually constructed it as follows. The tax data contain a classification of economic activity according to the (outdated) Dutch statistical classification of industries from 1974. At first glance, most webshops seemed to be classified as *retail trade*, many as *wholesale trade* and some as another type of industry according to the outdated classification. We constructed a training data set of 180 companies by manually categorising all companies of which the total annual turnover exceeded an industry-dependent threshold value in at least one year (see the second column of Table 5.2). The last column of Table 5.2 displays the number of companies manually categorised per type of industry. In fact, two manual categorisations (see Fig. 5.2) were made for each company that was included in the training data set presented in Table 5.2. The first categorisation is whether the company is economically active as a retail company according to the business register. We remark that the economic activity reported in the business register might be different from that found in the tax return data set. The second categorisation is whether the company is actually a webshop or not, based on manually searching the Internet.

Within the training data set, 76 webshops were identified. Their total turnover in 2016 was equal to €724,542,550. The validation data set is obtained from the training data set by applying stratified 5-fold cross-validation. This is described in more detail in Section 5.3.1.

TABLE 5.3: Number of companies per industry class (using the Dutch statistical classification of industries from 1974) included in the test data set.†

Industry (1974)	*Total frequency*	*Count*	*Webshop count*
Retail Trade	1,393	19	6
Wholesale Trade	3,329	20	1
Other	17,718	40	6
Total	22,440	79	13

† The frequency of each industry class in the tax data is included, as well as the number of identified webshops per industry class in the test data set.

5.2.3 Test data set

To assess the goodness of fit of a classification algorithm, we constructed a labelled *test data set* as follows. For *retail trade* and *wholesale trade*, 20 companies that were not in the training data set were randomly selected. (One duplicate retail company had to be removed.) For *other*, 40 companies were selected. The companies selected have been manually categorised, following the same approach as discussed for the companies in the training set. The results of the manual categorisation are shown in Table 5.3.

5.3 Methods

In this section we discuss the data-driven methods that were used to estimate cross-border Internet purchases within the EU by Dutch consumers. The methods, which can be applied to any other EU member state as well, are presented in three parts. In Section 5.3.1, the methods that were used to estimate the industry class for companies in the data set of tax returns are specified. Section 5.3.2 outlines how we accurately estimate webshop turnover. We do this by correcting for biases introduced by inaccuracies in the methods that were used to identify foreign webshops. In Section 5.3.3, we summarise our data-driven supply-side approach for measuring cross-border Internet purchases within the EU.

5.3.1 Estimating the industry class

The general set-up is as follows. Consider a population of n companies indexed by a set I. For a company $i \in I$, the industry class is denoted by $s_i \in H$. The set H consists of only two industry classes, namely webshops ($s_i = 1$) and other companies ($s_i = 0$). The total turnover from sales of goods as reported in the tax returns in year t by company i is denoted by $y_{i,t}$. In the tax data, $y_{i,t}$ is given for each $i \in I$ and each $t \in \{2014, 2015, 2016\}$. The goal is to estimate the annual turnover from sales of goods of the foreign webshops in the data set for each year t, given by

$$\sum_{i \in I} s_i y_{i,t}. \tag{5.1}$$

We assume that the classification s_i does not depend on t, while a company's economic activity might in reality change over time, for example when two companies merge. However, the specific case of merging companies is handled correctly, as the merged company will show up as a new company $i \in I$ in the data. Other causes for changes in economic activity are not corrected for. This might be refined in future work by determining the company's classification periodically (e.g., once a year).

The challenge of estimating expression (5.1) is that the industry classes s_i are not observed but must be estimated instead. We propose to estimate s_i in two different ways (in 5.3.1.1 by the business register and in 5.3.1.2 by websites) and combine the two estimates into a final estimate of industry class s_i (see Fig. 5.2). The combined estimate of s_i is used to evaluate expression (5.1).

5.3.1.1 Estimating the industry class by the business register

We assume that the sales of goods as reported in the tax returns filed by foreign companies registered (in the business register we use) as a retail company according to the NACE (Rev. 2) code are precisely the cross-border Internet purchases within the EU by the consumers of the EU member state under consideration. In other words, if a company $i \in I$ is registered as a retail company in the business register, we set the estimated industry class by the business register $\widehat{s_i}^{\mathrm{BR}}$ to 1. If not, we set $\widehat{s_i}^{\mathrm{BR}} = 0$.

The challenge is that we cannot simply look up a company $i \in I$ in the business register, as the tax data and the business register do not share a common unique identifier. The two data sets must be merged by matching legal company names. The following four issues then arise. First, the type of business entity might be registered differently in both data sets (e.g., LTD or LIMITED). Second, the name of a company might be spelled differently in both data sets (e.g., Muller or Mueller) and taking such differences into account (i.e., performing probabilistic record linkage) is computationally expensive. Third, we must choose which string distance metric to use for quantifying such differences numerically. Fourth, a threshold on the permitted number of spelling differences between names belonging to the same company must be determined. To overcome these four issues, we propose the following four-step approach (I—IV), where each step addresses the corresponding issue. Implementation details can be found in Appendix 5.A. We emphasise that only step I applies solely to *firm-level* record linkage. For other applications considering probabilistic record linkage, steps II-IV of our approach may directly be used.

Step I: stemming company names. First, as a preprocessing step, (1) non-alphanumeric characters are replaced by white spaces, (2) all leading, trailing and duplicate white spaces are removed, and (3) all characters are converted to lower case. Then, we remove the type of business entity (e.g., LTD) from the legal company name. For this, we may apply suffix stripping (or *stemming*) techniques, because the type of business entity comprises the end of a legal company name. Our data-driven stemming approach is inspired by Lovins (1968) and Porter (1980), where the latter is claimed to be "the most common algorithm for stemming English" (Manning, Raghavan & Schütze, 2008, p. 32). Finally, three values are stored for each company in the tax data and the business register. Taking the German company *Muller GmbH*, for example, the following three values are stored: `stem = 'muller'`, `suffix = 'gmbh'`, `suffix_class = 'LTD/DE'`. The *suffix class* indicates the type of company (using UK equivalents) and the company's EU member state of establishment.

Step II: locality-sensitive hashing. The variety of possible spelling differences in names of companies from the entire EU is huge, so manually formulating rules for matching tax data and the business register on company names is infeasible. To automate name-based record linkage we use *approximate string matching*; see, e.g., Cohen et al. (2003) for an overview. This entails measuring the distance between names (from now on, strings) viewed as elements of a (typically high dimensional) metric space. The approximate string match (according to a metric d) in the business register of a string s from the tax data would be the string t

in the business register that minimises $d(s,t)$ for t in the business register. The problem is that the value $d(s,t)$ must be computed for each pair (s,t), which is computationally expensive. In our case, $22{,}440 \times 6{,}996{,}468 \approx 1.57 \times 10^{11}$ values are computed, which may take up to several days on a regular machine, depending on which string distance metric is used.

An efficient and elegant approach is to use locality-sensitive hashing (LSH) schemes, which can be thought of as randomised dimensionality reduction preserving string distance. We shall use the famous LSH scheme MinHash (Broder, 1997), which is locality-sensitive for the Jaccard distance on character n-grams, or n-shingles (Leskovec, Rajaraman & Ullman, 2014, Chapter 3).

Although MinHash enables a faster approximate evaluation of the string distance metric, it does not yet reduce the number of evaluations that are required to match the two data sources. To achieve this, we apply the LSH Forest data structure (Bawa et al., 2005) on the results of MinHash applied to the business register. This data structure can then be queried to retrieve, for any string s (from the tax data) and any natural number m, the m approximately most similar strings in the input data set (the business register) according to any metric that induces a locality-sensitive hashing family (including MinHash). Choosing $m = 100$, the approximate string matching and locality-sensitive hashing techniques have reduced the number of evaluations from 1.57×10^{11} to $22{,}440 \times 100 = 2.24 \times 10^{6}$ (spending only about 80 min of wall-clock computation time on a regular machine).

Step III: combining string distance metrics. What remains, is to find the closest match in the remaining $m = 100$ companies from the business register for each company from the tax data, according to some string distance metric d. As we are not necessarily interested in the closest match itself, but simply in the binary outcome 'match'–'no match', we store only the minimum distance. For other applications where the match itself is of interest (e.g., in general record linkage problems), store the m-vector of distances and apply the remainder of our approach accordingly.

Two choices must be made in advance. First, some string distance metric must be selected. Second, some threshold value for $d(s,t)$ must be determined, above which we consider the approximate match a real match. In existing work on probabilistic record linkage of firm-level sources of data, these two choices are made manually, and typically somewhat arbitrarily. Therefore, the accuracy of the results will not be as high as possible, in general. To increase the accuracy, we

TABLE 5.4: Overview of the ten classification algorithms we consider, including whether we refer to it as a linear, nonlinear or ensemble algorithm.

Type	*Algorithm*
Linear	Logistic regression (LR)
	Linear discriminant analysis (LDA)
	Linear support vector classification (LinSVC)
Nonlinear	k-nearest neighbours (kNN)
	Multinomial naive Bayes (MNB)
	Quadratic discriminant analysis (QDA)
	Support vector classification with radial basis function kernel (RBFSVC)
Ensemble	Random forest (RF)
	Gradient boosting (GB)
	AdaBoost (AB)

propose the following general *data-driven* approach: consider multiple string distance metrics at once and let a machine learning algorithm determine the optimal threshold values. Details on which string distance metrics we have combined can be found in Appendix 5.A.

Step IV: machine learning. As mentioned in step III, we propose to use machine learning to find the optimal threshold values for string distance metrics (above which we consider an approximate match a real match). The aim is to find a classification algorithm $\widehat{s}_i^{\mathrm{BR}}$ that accurately predicts the industry class $s_i^{\mathrm{BR}} \in H$ (i.e., whether company $i \in I$ is registered in the business register as a retail company). The algorithm will use the eight-dimensional vectors of distances constructed in step III as the features (see Appendix 5.A). Recall that, for each company in the training set and the test set, the class s_i^{BR} was observed by manually searching the business register.

We propose the following data-driven approach to select a classification algorithm and corresponding algorithm parameter settings that are optimal in predicting s_i^{BR}. First, ten of the most commonly used classification algorithms are selected to be examined. We note that this selection is not exhaustive and it might be extended in future work. The ten classification algorithms that we consider are depicted in Table 5.4. We consider the linear classification algorithms logistic regression (LR), linear discriminant analysis (LDA) and linear support vector classification (LinSVC). The nonlinear algorithms implemented are k-nearest

TABLE 5.5: An overview of the parameter grids for the ten algorithms that we examined.†

Algorithm	*Parameter grid*
LR	penalty : $\{l1, l2\}$; C: $\{0.001, 0.01, 0.1, 1, 10\}$
LDA	*(no parameters)*
LinSVC	C: $\{0.001, 0.01, 0.1, 1, 10\}$
kNN	k: $\{1, 3, 5, \ldots, 39\}$
MNB	α: $\left\{10^{-10}, 0.01, 0.1, 1\right\}$
QDA	*(no parameters)*
RBFSVC	C: $\{0.01, 0.1, 1, 10, 100\}$; γ: $\{0.001, 0.01, 0.1, 1\}$
RF	n: $\{50, 100, 200, 500\}$; d: $\{1, 2, 3, \ldots, 8\}$
GB	n: $\{50, 100, 200, 500\}$; d: $\{1, 2, 3, \ldots, 8\}$; λ: $\{0.01, 0.1, 1\}$
AB	n: $\{50, 100, 200, 500\}$; d: $\{1, 2, 3, \ldots, 8\}$; λ: $\{0.01, 0.1, 1\}$

† In estimating the algorithms LR, LinSVC, RBFSVC, RF and AB, the two class-weighting schemes, uniform and balanced, were also included in the parameter grid. Consult the scikit-learn documentation for parameter specifications (`http://scikit-learn.org/`, version 0.19.1).

neighbours (kNN), multinomial naive Bayes (MNB), quadratic discriminant analysis (QDA) and support vector classification with radial basis function kernel (RBFSVC). Furthermore, we examine three ensemble algorithms, namely random forest (RF), gradient boosting (GB) and AdaBoost (AB). The details of the specifications of the classification algorithms can be found in, e.g., Witten, Frank, Hall and Pal (2017), Han, Kamber and Pei (2011), or Hastie et al. (2009). We used the Python library scikit-learn (`http://scikit-learn.org/`, version 0.19.1) to implement the ten classification algorithms, see also Pedregosa et al. (2011).

For each of the ten algorithms, a grid of parameter combinations to be examined was specified. These grids are depicted in Table 5.5. For precise parameter specifications we refer to the scikit-learn documentation (`http://scikit-learn.org/`, version 0.19.1).

For each algorithm and each parameter combination in the parameter grid, stratified 5-fold cross-validation is performed on the training data set. Cross-validation is used to prevent overfitting. The choice of using five folds is based on Breiman and Spector (1992). It might introduce more variance than choosing 10 or 20 folds (Kohavi, 1995). However, because of the small size of the training data set, choosing 10 or 20 folds might lead to unstable results. Therefore, we have chosen to use *stratified* 5-fold cross-validation in order to reduce the variance, as suggested by Kohavi (1995). Furthermore, we optimise parameter settings using mean F1-scores over the five folds. We prefer F1 over accuracy because of the low base rate of webshops in the entire data set. We do not use the common metric AUC to optimise parameter settings, as it is known to possibly mask poor performance when facing imbalanced data (Jeni, Cohn & De La Torre, 2013). As our data are in fact strongly imbalanced, because of the low base rate of webshops, it does not seem wise to use AUC as optimising metric. Moreover, optimising AUC does not, in general, imply optimising F1 (Davis & Goadrich, 2006). Thus, for each algorithm the parameter setting that maximises the mean F1-score over the five folds is selected. Subsequently, the mean and standard deviation of F1-scores over the five folds between the ten optimal classification algorithms are compared.

Finally, both the mean F1-score and the standard deviation of F1-scores over the five folds are considered in choosing the final classification algorithm and corresponding parameter settings. If necessary, the local behaviour on the parameter grid is examined to reduce the standard deviation of F1-scores over the five folds. This final classification algorithm is then trained on the entire training data set. The trained classification algorithm is used to compute the estimate $\widehat{s_i}^{\mathrm{BR}}$ for each company i not included in the training data set. Recall that $\widehat{s_i}^{\mathrm{BR}}$ is an estimate of s_i^{BR}, which indicates whether company i is registered as a retail company in the business register. In practice, it might be different from the true industry class s_i that we aim to estimate. For this reason, we propose estimating the industry class by websites as well, resulting in a second estimate $\widehat{s_i}^{\mathrm{W}}$.

5.3.1.2 Estimating the industry class by websites

We assume that a webshop can be identified by a shopping cart on the home page, referred to as such in the underlying HTML code. If a shopping cart is found on the website of company $i \in I$, we set the estimated industry class by websites $\widehat{s_i}^{\mathrm{W}}$ to 1. If not, we set $\widehat{s_i}^{\mathrm{W}} = 0$. For this classification, we propose the following three-step approach. First, as the tax data do not contain the URL of the website of a company, we implement a method for finding this URL based on the legal company name. Second, web scraping is used to look for a shopping cart on the website. Third, as the first two steps are not flawless, the machine learning approach from Section 5.3.1.1 (step IV) is used to minimise errors. The following three parts describe these three steps in more detail.

Step I: finding a company's website. In tax data from the Netherlands, a URL of the home page of the company is not available. Therefore, Statistics Netherlands has developed *URL retrieval software* to retrieve the URL of the home page of a company based on the legal company name (Ten Bosch & Windmeijer, 2018). The legal company name is first processed by Google's search application programming interface, returning a list of several URLs, each equipped with a title and a description. The URLs are then ranked according to a *matching score* between 0 (definitely not a match) and 1 (definitely a match), which is computed by a random-forest algorithm. The algorithm is trained by using a set containing 1000 Dutch company names (from different industries and varying in size, i.e., the number of employees) and the URL of their website. We emphasise that the Dutch language of the training set is not necessarily an issue, as most foreign webshops selling to Dutch consumers will have a Dutch version of the website (see Section 5.1). For each company, the URL with the highest assigned matching score is returned and the corresponding matching score is stored.

Step II searching for a shopping cart. For each company, the HTML code of the URL found in step I is downloaded as a raw text file. In the raw text file, the occurrences of variations of the words *shop* and *cart* in Dutch, English and German are counted. The full list is *winkel, wagen, mand, shop, cart, bag, basket, warenkorb*. The choice of these three languages is based on the fact that most Dutch citizens mostly speak only Dutch, English and/or German. Note that in modern information retrieval, it is more common to count the occurrences of all words found in a document (see Manning et al., 2008, Chapter 6). We have chosen not to follow this approach, as it would lead to serious dimensionality issues; the number of different terms (words) would be much larger than the number of documents (websites of companies) in the training set.

Step III: machine learning. The final step is to find a classification algorithm $\widehat{s}_i^{\text{W}}$ that can accurately predict the industry class $s_i \in H$. The true, unobserved industry class $s_i^{\text{W}} \in H$ represents whether company $i \in I$ is a webshop. Recall that, for each company in both the training and test set, the class s_i^{W} was observed by manually searching the Internet.

Before training a classification algorithm, the counts of the words are transformed to real numbers in the interval $[0, 1]$ by using (normalised) term-frequency times inverse-document-frequency (TF.IDF) (see Witten et al., 2017, p. 314, for a definition). To prevent division by 0 in computing the IDF, a single document containing each of the eight words once is added to the data. The eight TF.IDF values and the maximum matching score are used as features in fitting classification algorithms on the training data. The machine learning approach is identical to that described in step IV of Section 5.3.1.1.

5.3.1.3 Constructing the final estimate of the industry class

The two selected classification algorithms, each with the optimal parameter setting, are trained on the entire training data set. The trained models are used to compute $\widehat{s_i}^{\mathrm{BR}}$ and $\widehat{s_i}^{\mathrm{W}}$ on the remaining part of the data set. Companies whose features, which are needed for one of the two algorithms, are (partially) missing receive the value -1 as prediction, to be interpreted as 'missing'. It happens for $\widehat{s_i}^{\mathrm{BR}}$ if the tax-stem of the company has less than three characters. It happens for $\widehat{s_i}^{\mathrm{W}}$ if the maximum matching score is below 0.5 or if no HTML code was downloaded. The final single categorisation $\widehat{s_i}$ is obtained by combining $\widehat{s_i}^{\mathrm{BR}}$ and $\widehat{s_i}^{\mathrm{W}}$ as follows:

$$\widehat{s_i} := \begin{cases} -1 & \text{when } \widehat{s_i}^{\mathrm{BR}} = \widehat{s_i}^{\mathrm{W}} = -1, \\ \widehat{s_i}^{\mathrm{BR}} & \text{when } \widehat{s_i}^{\mathrm{W}} = -1, \\ \widehat{s_i}^{\mathrm{W}} & \text{when } \widehat{s_i}^{\mathrm{BR}} = -1, \\ \widehat{s_i}^{\mathrm{BR}} \wedge \widehat{s_i}^{\mathrm{W}} & \text{otherwise.} \end{cases}$$

The AND-operator $\wedge$ is computed as the minimum of the two integers. It implies that $\widehat{s_i}$ categorises a company as a webshop if and only if the company is categorised as such by both $\widehat{s_i}^{\mathrm{BR}}$ and $\widehat{s_i}^{\mathrm{W}}$ (see Fig. 5.2).

5.3.2 Accurately estimating webshop turnover

Estimating webshop turnover once $\widehat{s_i}$ has been estimated seems straightforward: simply use it instead of s_i to evaluate expression (5.1). However, this straightforward evaluation will result in a biased estimation of webshop turnover. This section aims to estimate and correct that bias, yielding a more accurate estimate of webshop turnover.

5.3.2.1 Biased estimation of webshop turnover

We begin by isolating the bias in estimating expression (5.1). For companies in the training set, the manual categorisation s_i is the true class of company $i \in I$. Hence, rewriting expression (5.1), the total (annual) cross-border Internet purchases could thus be estimated as

$$\sum_{i \in I_M} s_i y_i + \sum_{i \in I \backslash I_M} \widehat{s}_i y_i, \tag{5.2}$$

where $I_M \subset I$ is the training set of manually categorised companies. The first term is the total turnover of observed webshops in the training data set and the second term is the total turnover of predicted webshops in the rest of the data set.

Now, Fig. 5.3 illustrates that expression (5.2) yields (because of the second term) a biased estimate of expression (5.1). In fact, any aggregate based on the results of a classification algorithm will be a biased estimate of the true value. For a binary classifier, the only exception is when the number of false positive predictions is equal to the number of false negative predictions, which is equivalent to precision and recall being equal. To the best of our knowledge, we are the first to note this in the setting of machine learning.[4]

Before estimating and correcting the bias we introduce the vector notation from Van Delden et al. (2016). We write $\boldsymbol{a_i}$ for the 2-vector $(s_i, 1 - s_i)^T$ and consider the aggregate turnover vector $\boldsymbol{y} = \sum_{i \in I} \boldsymbol{a_i} y_i$. Similarly, define $\widehat{\boldsymbol{a}}_{\boldsymbol{i}}$ based on $\widehat{s}_i$. Expression (5.2) will thus become the first component of the estimated 2-vector $\widehat{\boldsymbol{y}}$ given by

$$\widehat{\boldsymbol{y}} := \sum_{i \in I_M} \boldsymbol{a_i} y_i + \sum_{i \in I \backslash I_M} \widehat{\boldsymbol{a}}_{\boldsymbol{i}} y_i. \tag{5.3}$$

In the remainder of this section, only the subset $I \backslash I_M \subset I$ is considered. Hence, any index i will refer to a company that is not in the training set I_M. Consequently, the estimate $\widehat{\boldsymbol{y}}$ will be used to refer only to the second term on the right-hand side of equation (5.3), as the first term does not introduce any bias. Similarly, $\boldsymbol{y}$ will be used in the remainder of this section to refer to $\sum_{i \in I \backslash I_M} \boldsymbol{a_i} y_i$.

5.3.2.2 Classification-error model

To estimate and correct the bias, we follow the approach of Van Delden et al. (2016), which did not focus on machine learning algorithms, but it can directly be applied in that setting, as we will show below. The approach entails that s_i is considered to be deterministic and $\widehat{s}_i$ to be stochastic, conditionally on s_i. They

[4] As mentioned in Section 5.1, we were alerted to the literature on quantification learning only after publication of this chapter. It was first noted in the setting of machine learning by Forman (2005).

assume the following *classification-error model*

$$p_{ghi} := \mathbb{P}(\widehat{s}_i = h \mid s_i = g), \qquad g, h \in H. \tag{5.4}$$

We emphasise that this assumption is very reasonable if $\widehat{s}_i$ is the result of a machine learning algorithm. Such an algorithm is mostly based on assuming a data-generating process, where the independent variable s_i is assumed to be a function of dependent variables (or features) that result in $\widehat{s}_i$, plus an error or noise term. This noise term corresponds to the stochastic classification error model above.

In addition, we assume that p_{ghi} does not depend on $i \in I \backslash I_M$. This assumption might be argued to be incorrect for two reasons. First, it is more difficult to find the correct website for a small company than for a large company. Moreover, a small company that is not a webshop might not even have a website. Second, the coverage and quality of the business register for smaller companies is significantly lower than for larger companies. Both reasons imply that the probability of a classification error (more specifically, a false negative classification error) increases as turnover decreases. However, we make the assumption because accurately estimating P for different turnover classes, as suggested by Van Delden et al. (2016), requires a far larger training data set than the training data set that we have available.

The resulting 2×2 matrix $P = (p_{gh})_{g,h \in H}$ is estimated as follows. On the test data set, $\widehat{s}_i$ is compared to s_i. Denoting by TP, FP, TN, and FN the number of true and false positive and true and false negative classifications respectively, the estimator $\widehat{P}$ of P takes the form

$$\widehat{P} = \begin{pmatrix} \dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} & \dfrac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \\[2ex] \dfrac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} & \dfrac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \end{pmatrix}. \tag{5.5}$$

We next show how to use this estimator to obtain accurate estimates of cross-border Internet purchases.

5.3.2.3 Estimating bias and variance

If we assume the classification-error model, it follows $\mathbb{E}(\widehat{\boldsymbol{a}}_{\boldsymbol{i}}) = P^T \boldsymbol{a}_{\boldsymbol{i}}$ and therefore

$$\mathbb{E}(\widehat{\boldsymbol{y}}) = P^T \boldsymbol{y}. \tag{5.6}$$

The bias of $\widehat{\boldsymbol{y}}$ as an estimator of $\boldsymbol{y}$ equals

$$\boldsymbol{B}(\widehat{\boldsymbol{y}}) = \mathbb{E}(\widehat{\boldsymbol{y}}) - \boldsymbol{y} = (P^T - I_2)\boldsymbol{y}, \tag{5.7}$$

where I_2 is the 2×2 identity matrix. This shows that, in general, expression (5.3) yields a biased estimate of $\boldsymbol{y}$. In fact, the bias is only zero if either (1) the classification algorithm does not make any errors (i.e., $P^T = I_2$) or (2) $\boldsymbol{y}$ precisely equals an eigenvector of P^T corresponding to the eigenvalue 1.

To estimate the bias as given in expression (5.7), we could use the plug-in estimator

$$\widehat{\boldsymbol{B}}_{\boldsymbol{0}} = (\widehat{P}^T - I_2)\widehat{\boldsymbol{y}}. \tag{5.8}$$

Following Van Delden et al. (2016), we assume that $\mathbb{E}[\widehat{P}^T] = P^T$ and that $\widehat{P}^T$ and $\widehat{\boldsymbol{y}}$ are uncorrelated. It follows that $\mathbb{E}[\widehat{\boldsymbol{B}}_{\boldsymbol{0}}] = P^T \boldsymbol{B}(\widehat{\boldsymbol{y}})$, hence the plug-in estimator is a biased estimator of the bias. If we assume that $p_{01} + p_{10} \neq 1$ (and $\widehat{p}_{01} + \widehat{p}_{10} \neq 1$), then the inverse matrix $Q = (P^T)^{-1}$ exists (and $\widehat{Q} = (\widehat{P}^T)^{-1}$ exists). Now, assuming $\mathbb{E}[\widehat{Q}] = Q$ and that $\widehat{Q}$ and $\widehat{y}$ are uncorrelated, an unbiased estimator of the bias is

$$\widehat{\boldsymbol{B}}_{\boldsymbol{1}} = (I_2 - \widehat{Q})\widehat{\boldsymbol{y}}. \tag{5.9}$$

However, correcting $\widehat{\boldsymbol{y}}$ by $\widehat{\boldsymbol{B}}_{\boldsymbol{1}}$ might increase the variance of (the first component of) the estimator. It might lead to low accuracy in practice. Therefore, Van Delden et al. (2016) propose to find the optimal value $\lambda = \lambda^*$ for which the MSE of the first component of $\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}} = (1-\lambda)\widehat{\boldsymbol{B}}_{\boldsymbol{0}} + \lambda\widehat{\boldsymbol{B}}_{\boldsymbol{1}}$ as an estimator of the bias $\boldsymbol{B}(\widehat{\boldsymbol{y}})$ is minimised for $\lambda \in [0, 1]$. For more details on how to derive λ^*, please consult Appendix 5.B.

Having found λ^*, we estimate $\boldsymbol{y}$ (still excluding $\boldsymbol{y}_{\boldsymbol{M}}$) by the first component of the vector

$$\begin{aligned}\widehat{\boldsymbol{y}}_{\lambda^*} &= \widehat{\boldsymbol{y}} - \widehat{\boldsymbol{B}}_{\lambda^*} = \widehat{\boldsymbol{y}} - (I_2 + \lambda^*(\widehat{Q} - I_2))(\widehat{P}^T - I_2)\widehat{\boldsymbol{y}} \\ &= \left(2I_2 - \widehat{P}^T - \lambda^*(\widehat{Q} - I_2)(\widehat{P}^T - I_2)\right)\widehat{\boldsymbol{y}}.\end{aligned} \tag{5.10}$$

The standard deviation is estimated by the square root of the upper-left value in the variance-covariance matrix

$$\widehat{V}(\widehat{\boldsymbol{y}}_{\Lambda^*}) = \left(2I_2 - \widehat{P}^T - \lambda^*(\widehat{Q} - I_2)(\widehat{P}^T - I_2)\right) \widehat{V}(\widehat{\boldsymbol{y}}) \left(2I_2 - \widehat{P}^T - \lambda^*(\widehat{Q} - I_2)(\widehat{P}^T - I_2)\right)^T . \tag{5.11}$$

Here, the variance $V(\widehat{\boldsymbol{y}})$ of $\widehat{\boldsymbol{y}}$ is estimated by

$$\widehat{V}(\widehat{\boldsymbol{y}}) = \mathrm{diag}\left(\widehat{P}^T\widehat{\boldsymbol{k}}\right) - \widehat{P}^T\mathrm{diag}\left(\widehat{\boldsymbol{k}}\right)\widehat{P}, \tag{5.12}$$

where $\widehat{\boldsymbol{k}} = \sum_i \widehat{\boldsymbol{a}}_i y_i^2$. The bias of $\widehat{V}(\widehat{\boldsymbol{y}})$ as estimator of $V(\widehat{\boldsymbol{y}})$ is relatively small and therefore is not corrected (Van Delden, Scholtus & Burger, 2015, Appendix A4). As the values of $\boldsymbol{y}_M$ are not stochastic, expression (5.11) also yields the standard deviation of the final estimate of $\boldsymbol{y}$.

5.3.3 Summarising our data-driven supply-side approach

The proposed data-driven supply-side approach for measuring cross-border Internet purchases within the EU can be summarised as follows. Based on EU VAT legislation, the starting point is a data set of tax returns filed by foreign companies established within the EU. These tax data are *supply-side* data as they contains company turnover. Then, the challenge is to identify webshops within the data set of tax returns. We address this challenge in two steps. In the first step, we implement approximate string matching techniques to merge the tax data to a business register of retail companies that are established within the EU. The merging can be viewed as *data-driven* record linkage, as we optimised the performance of the approximate string matching by using machine learning algorithms. In the second step, we use web scraping in combination with machine learning to assess whether a company is a webshop. The outcomes of the two steps are combined to obtain a more accurate estimate of cross-border Internet purchases. Moreover, we use the data to estimate the bias and standard deviation of the estimate. Thus, the data-driven methods applied to the supply-side data yield our data-driven supply-side approach for measuring cross-border Internet purchases within the EU.

5.4 Results

Below, we present our findings of applying the approach to the Netherlands by estimating cross-border Internet purchases within the EU by Dutch consumers. The section is structured as follows. First, in Section 5.4.1, the results of training the classification algorithms to estimate the industry class by the business register are presented. Then, in Section 5.4.2, the same is presented for estimating the industry class by websites. Next, in Section 5.4.3, we present the results of estimating cross-border Internet purchases by Dutch consumers. It contains the most relevant results of the paper. Finally, in Section 5.4.4, we compare the results of our data-driven supply-side approach to currently available results from demand-side approaches based on consumer surveys. We interpret and discuss the differences of the resulting estimates of cross-border Internet purchases by Dutch consumers.

5.4.1 Results from estimating the industry class by the business register

As can be seen from the results in Table 5.6, machine learning is very well suited for probabilistic record linkage of firm-level data. Recall from step IV of Section 5.3.1.1 that we compared ten different machine learning algorithms (Table 5.4), each evaluated by using multiple parameter settings (Table 5.5), to estimate the industry class s_i^{BR} by the business register. Table 5.6 does not include results for MNB; this algorithm assumes discrete features, whereas they are continuous (distances between strings). For each algorithm, we have selected the parameter settings that are optimal in estimating s_i^{BR}, based on the mean F1-score from the stratified 5-fold cross-validation. The results in Table 5.6 show that the mean goodness of fit of the machine learning algorithms are high, with little difference between the algorithms. Moreover, the standard deviations in scores over the folds (shown in parentheses) are small.

The final classification algorithm that we use to predict s_i^{BR} is RBFSVC, with parameters $C = 100$, $\gamma = 1$ and the balanced class-weighting scheme (see Table 5.6). Observe that this choice not only maximises the mean F1-score, but also mean precision and mean recall. In particular, the algorithm does not falsely predict positive classifications on the training data set. Moreover, the local behaviour of the mean F1-score of RBFSVC, as a function of the parameters C and γ, is stable around the optimal parameters (see Appendix 5.C).

TABLE 5.6: Mean (plus or minus standard deviation) of scores for optimal parameter settings for each of the specified algorithms estimating s_i^{BR}.[†]

Algorithm	*Optimal parameters*	*F1* ↓	*Precision*	*Recall*
RBFSVC	$C = 100, \gamma = 1$	**0.97 (± 0.03)**	**1.00 (± 0.00)**	**0.94 (± 0.05)**
GB	$n = 50, d = 1, \lambda = 0.01$	0.95 (± 0.02)	0.98 (± 0.03)	0.92 (± 0.03)
kNN	$k = 3$	0.95 (± 0.03)	0.98 (± 0.03)	0.93 (± 0.04)
LinSVC	$C = 0.01$	0.94 (± 0.02)	0.97 (± 0.03)	0.91 (± 0.03)
LDA		0.94 (± 0.03)	**1.00 (± 0.00)**	0.89 (± 0.05)
LR	$C = 1$, L1-penalty	0.94 (± 0.03)	0.97 (± 0.03)	0.92 (± 0.03)
AB	$n = 100, d = 1, \lambda = 0.1$	0.94 (± 0.04)	0.96 (± 0.04)	0.93 (± 0.04)
RF	$n = 50, d = 4$	0.94 (± 0.04)	0.95 (± 0.04)	0.93 (± 0.04)
QDA		0.93 (± 0.02)	**1.00 (± 0.00)**	0.87 (± 0.03)

[†] The scoring function F1 (used to rank the results) is used to optimise across the parameter settings in the parameter grid. Each parameter setting is evaluated using stratified 5-fold cross-validation. In each column, the maximum score is highlighted. In the fourth column three scores are the maximum score (rows RBFSVC, LDA, and QDA).

5.4.2 Results from estimating the industry class by websites

The results in Table 5.7 show lower scores and greater difference between algorithms than the results in Table 5.6. Again, the algorithms (which are now used to estimate the industry class s_i^{W} by websites) are ranked with respect to the optimal mean F1-score over the folds in the stratified 5-fold cross-validation. Also, note that the standard deviations of scores over the folds (which are shown in parentheses) are relatively large. Analysing the results more closely, using the (simple) categorisation of the machine learning algorithms that were used into linear, non-linear and ensemble algorithms (see Table 5.4), we make an interesting observation. Based on the results in Table 5.7, all three algorithms from the category of linear methods (LR, LinSVC, LDA) performs less well than the (better performing) algorithms from the other two categories of methods. This suggests that a linear separation of the data points in higher dimensional space does not yield the best classification for unseen data. Therefore, it could be more difficult to estimate the industry class by websites than by the business register, leading to the considerable differences between the results of the two estimations. Hence, in future work it might be worthwhile to obtain more training data to improve the results.

The final classification algorithm that we use to estimate s_i^{W} is RF, with parameters $n = 200$, $d = 1$ and the balanced class-weighting scheme (see Table 5.7). The

Table 5.7: Mean (plus or minus standard deviation) of scores for optimal parameter settings for each of the specified algorithms predicting s_i^{W}.†

Algorithm	*Optimal parameters*	*F1* ↓	*Precision*	*Recall*
AB	$n = 100$, $d = 1$, $\lambda = 0.1$, bal.	**0.80 (± 0.11)**	0.82 (± 0.10)	0.78 (± 0.12)
GB	$n = 200$, $d = 1$, $\lambda = 0.1$	0.79 (± 0.10)	0.80 (± 0.09)	0.78 (± 0.12)
RF	$n = 200$, $d = 1$, bal.	0.78 (± 0.10)	**0.85 (± 0.14)**	0.76 (± 0.16)
kNN	$k = 35$	0.76 (± 0.09)	0.81 (± 0.04)	0.73 (± 0.17)
RBFSVC	$C = 1$, $\gamma = 0.1$, bal.	0.76 (± 0.11)	0.78 (± 0.08)	0.76 (± 0.18)
LR	$C = 1$, L1-penalty	0.75 (± 0.12)	0.76 (± 0.10)	0.76 (± 0.18)
LinSVC	$C = 0.01$	0.74 (± 0.10)	0.77 (± 0.07)	0.73 (± 0.17)
LDA		0.74 (± 0.13)	0.69 (± 0.14)	**0.81 (± 0.17)**
MNB	$\alpha = 10^{-10}$	0.70 (± 0.12)	0.71 (± 0.15)	0.71 (± 0.12)
QDA		0.67 (± 0.15)	0.63 (± 0.12)	0.73 (± 0.21)

† The scoring function F1 (used to rank the results) is used to optimise across the parameter settings in the parameter grid. Each parameter setting is evaluated using stratified 5-fold cross-validation. In each column, the maximum score is highlighted.

reason for this choice is that RF maximises mean precision. Moreover, the local behaviour of the F1-score of RF, as a function of the algorithm parameters, is more stable around the optimal parameters compared with the local behaviour for AB and GB (see Appendix 5.C).

5.4.3 Estimating cross-border Internet purchases

In Table 5.8, we present our final estimates of cross-border Internet purchases within the EU by Dutch consumers. Recall that the algorithm chosen in Section 5.4.1 has now been (re)trained on the entire training data set indexed by I_M. It resulted in a model $\widehat{s}_i^{\mathrm{BR}}$ that was qualified to predict s_i on the remaining part of the data set of tax returns, indexed by $I \backslash I_M$. Similarly, a model $\widehat{s}_i^{\mathrm{W}}$ has been trained by using the algorithm chosen in Section 5.4.2. The two models were combined into a final model $\widehat{s}_i$ (as described in Section 5.3.1.3, see also Fig. 5.2). The comparison between the model $\widehat{s}_i$ and the observed true values s_i on the test data set yields the values TP = 8, FP = 4, TN = 62 and FN = 5. It follows that

$$\widehat{P} = \begin{pmatrix} 8/13 & 5/13 \\ 4/66 & 62/66 \end{pmatrix} \approx \begin{pmatrix} 0.615 & 0.385 \\ 0.061 & 0.939 \end{pmatrix}.$$

The main results of the paper are shown in Table 5.8. The values y_M contain the total cross-border Internet purchases at companies in the set I_M. The categorisation for companies in I_M has been determined manually and can be considered free from errors. The values $\widehat{y}$ contain the additional estimated cross-border Internet purchases at companies in the set $I \backslash I_M$. The values λ_{opt} contain the optimal values of λ in minimising the MSE of the estimated bias of $\widehat{y}$. Note that all optimal values of λ are equal to 0, meaning that the increased variance dominates the decreased squared bias of $\widehat{\boldsymbol{B}}_{\boldsymbol{1}}$ compared to $\widehat{\boldsymbol{B}}_{\boldsymbol{0}}$. This is due to the relatively large off-diagonal values in the matrix $\widehat{P}$. The values $\widehat{B}_{\lambda_{\text{opt}}}$ represent the estimated bias of $\widehat{y}$ for the optimal value $\lambda = \lambda_{\text{opt}}$. Note that the bias strongly differs across the three years. The values y show the final estimate of the total cross-border Internet purchases, computed as

$$y = y_M + (\widehat{y} - \widehat{B}_{\lambda_{\text{opt}}}). \tag{5.13}$$

The last column in Table 5.8 contains the standard deviation of y, estimated as outlined at the end of Section 5.3.2.

In the Netherlands, total household consumption on retail goods (food and durable goods, codes 1000 up until and including 3000) in 2016 was equal to €87,206 million, according to Statistics Netherlands (`https://opendata.cbs.nl`). Statistics Netherlands does not publish the total on-line consumption of goods by Dutch consumers. The only currently available estimate is by Thuiswinkel.org and GfK and it is based on consumer surveys. The estimate of 2016 equals €11.01 billion. It seems possible that just over 12% of on-line consumption by Dutch consumers is spent at foreign webshops established within the EU. Besides, Statistics Netherlands does publish year-on-year growth figures on on-line retail sales by Dutch webshops. In 2016, this year-on-year growth was equal to 22.1%. It is quite similar to the growth of 21.2% that we find by comparing the values of y in 2015 and 2016 as presented in Table 5.8.

Reflecting on our findings, we note that the standard deviation of the final estimate would still be too large for official statistical purposes. However, as will be discussed more thoroughly in Section 5.4.4, our findings prove to be a significant improvement over currently available alternative estimates.

TABLE 5.8: Final results of cross-border Internet purchases within the EU by Dutch consumers in millions of euros.

Year	y_M	$\widehat{y}$	λ_{opt}	$\widehat{B}_{\lambda_{\text{opt}}}$	y	*Std*(y)
2014	€ 405 M	€ 495 M	0	€ 63 M	€ 837 M	€ 97 M
2015	€ 565 M	€ 586 M	0	€ 21 M	€ 1,132 M	€ 101 M
2016	€ 725 M	€ 667 M	0	€ 19 M	€ 1,372 M	€ 110 M

5.4.4 Comparison with demand-side approach

In Section 5.1, we claimed that our data-driven supply-side approach would be more accurate than demand-side approaches to estimate cross-border Internet purchases within the EU. To justify this claim, we compare our results for the Netherlands to the results of the consumer survey approach by market research institute GfK (commissioned by Thuiswinkel.org on behalf of Ecommerce Europe). We choose to use the estimate by these commercial organisations, as, to the best of our knowledge, there is no scientific literature reporting the total cross-border Internet purchases by Dutch consumers.

In 2016, total cross-border on-line consumption by Dutch consumers according to GfK was equal to €637 million, €190 million of which were spent in China and €70 million in the United States. This implies that at most €377 million were spent within the EU, but this figure includes on-line consumption of both goods and services.

Moreover, the fraction of on-line consumption of goods in the total on-line consumption in 2016, as reported by GfK, was €11.01 billion / €20.16 billion = 0.55. We assume that this proportion is independent of the country in which the goods or services were purchased. As a result, cross-border on-line purchases of goods within the EU, according to GfK, would approximately equal €206 million in 2016.

We, however, find €1,372 million for 2016 with a standard deviation of €110 million, i.e., 8%. The estimate is more than six times as high as that of GfK. The results show the severe downward bias in using demand-side approaches to estimate cross-border on-line consumption and it motivates the implementation of our approach for other EU member states.

5.5 Chapter conclusions

We have proposed a methodology to measure cross-border Internet purchases within the EU by using tax data, a business register, and website data. We have implemented data-driven methods to combine these supply-side data sources in a computationally efficient manner. Applied to the Netherlands, the proposed approach leads to a strong improvement of existing approaches that are based on consumer surveys. In particular, market research institute GfK (commissioned by Thuiswinkel.org on behalf of Ecommerce Europe) use consumer surveys and estimated cross-border Internet purchases by Dutch consumers within the EU in 2016 to be approximately €206 million. Our approach yields an estimate of €1,372 million, i.e., six times as high as GfK's estimate, with a standard deviation of €110 million (8%).

The approach that we propose requires foreign companies' tax returns to contain only the legal company name and the turnover from sales of goods to consumers. Because of EU VAT legislation these data are available in every EU member state. In fact, we do not require the economic activity of a company to be accurately available in filed tax returns. The training and test set could even be constructed without any known economic activity, by viewing all companies as belonging to the same class and following the construction described in Sections 5.2.2 and 5.2.3. We also do not assume that the URL of the home page of a company is available in filed tax returns. Moreover, the additional data (the business register and websites) required by the approach proposed are open data sources. Hence, our main conclusion is that the approach is applicable in any EU member state and more accurately estimates cross-border Internet purchases within the EU.

In addition to our methodological contribution to official statistics concerning cross-border Internet purchases, we point out two aspects of our contribution that might be of interest to a general audience in statistics. The first aspect is our fully data-driven (and therefore generic) approach for probabilistic record linkage of firm-level data sources. The novelty of our approach compared with the extant literature is that we use machine learning to maximise the accuracy of the record linkage. In this regard, we improve upon existing methods as the optimisations (choosing an optimal string matching algorithm and similarity threshold) are fully automated. Moreover, our approach is computationally efficient by using state-of-the-art hashing techniques from computer science. Therefore, our approach can be applied to related large-scale (text-based) probabilistic record linkage problems. The second aspect is the observation that aggregation (e.g., summing) after running a classification algorithm yields (potentially strongly) biased estimates. We believe that we are the first to make this observation in the field of machine learning. As a first step, we have shown

(1) that, in many fields outside machine learning, techniques have been developed to correct the bias of aggregate estimates and

(2) that we can directly apply the techniques to classification algorithms.

In our view, the bias of aggregates based on results from classification algorithms deserves more investigation beyond our first step.

Although our new methodology improves the estimation of cross-border Internet purchases within the EU, we point out two potential sources of bias of our current approach. First, companies with sales below the threshold value in the country of destination (in the Netherlands: €100,000) do not have to file a tax return. The Internet purchases at such small companies are therefore missing in an estimation based on tax data. Yet, the sales of small companies via marketplaces (e.g., Amazon) are included in tax data. Second, the reported turnover from sales to consumers might be inaccurate, potentially leading to an underestimation of total cross-border Internet purchases. However, this underestimation is expected to be minimal because of strict law enforcement by, and collaboration between, tax authorities in the EU. We have not aimed to correct for either of these two biases, as no data are available to estimate them. Moreover, we aimed to show the downward bias of consumer survey approaches compared with a supply-side approach in estimating cross-border Internet purchases within the EU. We therefore do not mind if our supply-side approach still yields a conservative estimate.

Future work on measuring cross-border Internet purchases within the EU might focus on improving the predictions by websites of company classifications. The empirical results show that this is the weakest part of the approach that we propose, as the F1-scores for website-based predictions are lower than the F1-scores for the predictions based on using the business register. The results may be improved by enlarging the training set of the URL retrieval software from Dutch to European websites using the company names and URLs registered in the business register. We consider this improvement outside the scope of the current paper, as the results of our data-driven supply-side approach already show a strong improvement compared with existing consumer survey approaches.

Finally, further applications of the data-driven supply-side approach include revealing the structure of the cross-border on-line retail market in any EU member state. Our approach directly returns a list of foreign webshops and their annual cross-border Internet sales to the observing member state. If the information on domestic webshops that are active within the member state's e-commerce market is complemented, the structure of that market may be analysed. Related to this is the export of the webshops established in a single EU member state, being the supply-side counterpart of cross-border on-line consumption within a member state. It might be interesting to compare the two market structures within individual member states and to compare the market structures between member states within the EU.

APPENDIX

5.A Estimating the industry class by the business register

This appendix contains the details of the first three steps of our four-step approach for estimating the industry class by the business register. The details of step IV can be found in the main text.

Step I: stemming company names

The stemming of company names in the tax data and the business register follows the following three steps, inspired by Lovins (1968) and Porter (1980).

- *Step 1:* use the business register to create, for each `country` in the EU and each `n = 1,2,3,4`, a list of the five most common legal company name *suffixes* (i.e., end-of-string words) of length `n`. Complement the list with the types of business entities from Table 5.9.

- *Step 2:* for each EU member state, identify the *prefixes* of the suffixes in the list obtained in step 1 as well as the *suffix-class* (of the form 'business type/member state'). Concatenate the suffix-prefix lists obtained into a single list.

- *Step 3:* for each legal company name, search each of its words (starting from the second word) in the suffix-prefix list from step 2. Stop if a match is found. Split the name into a *stem* and a suffix, storing its suffix class.

We include an example to illustrate the stemming procedure. In Germany, the most common type of business entity is *Gesellschaft mit beschränkter Haftung* (GmbH), which is similar to a private company limited by shares (LTD).

Table 5.9: Overview of the types of business entities per EU member state, obtained from Wikipedia (https://en.wikipedia.org/wiki/List_of_business_entities).

Country	*Code*	Types of business entities
Austria	AT	AG, GmbH, KG, GmbH & Co. KG
Belgium	BE	BVBA, NV, SA
Bulgaria	BG	AD, EAD, EOOD, OOD
Croatia	HR	d.d., d.o.o.
Cyprus	CY	(Same as GB)
Czech Republic	CZ	a.s., s.r.o.
Denmark	DK	ApS, A/S, A.M.B.A.
Estonia	EE	OÜ, AS
Finland	FI	Oy, oyj
France	FR	SARL, SA
Germany	DE	OHG, KG, AG, GmbH, GmbH & Co. KG/AG/OHG
Greece	GR	A.E., E.P.E.
Hungary	HU	*korlatolt felelossegu tarsasag*, *reszvenytarsasag*
Ireland	IE	(same as GB)
Italy	IT	s.r.l., s.p.a., *societa a responsabilita limitata*
Latvia	LV	SIA, AS
Lithuania	LT	UAB, AB
Luxembourg	LU	S.A., S.A.R.L., SECS
Malta	MT	(same as GB)
Netherlands	NL	BV, NV
Poland	PL	Sp. Z.O.O., S.A.
Portugal	PT	lda., S.A.
Romania	RO	S.R.L., S.A.
Slovakia	SK	S.R.O., A.S.
Slovenia	SI	d.d., d.o.o.
Spain	ES	S.A, *sociedad anonima*, S.L., *sociedad limitada*
Sweden	SE	AB, *aktiebolag*
United Kingdom	GB	*private limited company*, ltd, *limited*, plc, public limited company

This type of business entity will show up in step 1 for `country = ‘DE’` and `n = 4`. Many variations may occur due to partial abbreviations (e.g., *Gesellschaft mbH*). Step 2 ensures that only three suffix-prefixes must be searched for in step 3: GMBH, G M B H, and Gesellschaft. The *suffix-class* corresponding to each of these three suffix-prefixes is ‘private company limited by shares, German’ (LTD/DE). Now, take as an example a German company named *Muller GmbH*, which is stored as `muller gmbh` after the preprocessing step. The algorithm starts searching the second word, `gmbh`, in the suffix-prefix list. It is found, and three values are stored for this company: `stem = ‘muller’`, `suffix = ‘gmbh’`, `suffix_class = ‘LTD/DE’`. In general, if the second word is not found, the algorithm would continue with the third word, until the name's final word. If no matches are found at all, the stem equals the name and we obtain `suffix = ‘’` and `suffix_class = ‘’` (empty strings).

Step II: locality-sensitive hashing

The tax data and the business register do not contain characters with diacritical marks (e.g., the German umlaut as in 'ü'). A German name such as Müller (English: Miller) has been registered using plain alphabetic characters instead. For the 'ü' in Müller, two conventions exist: Muller or Mueller. This leads to potential spelling differences between the tax data and the business register for the same company. Another common difference is the use or omission of white spaces (e.g., webshop *versus* web shop).

As discussed in the main text, we use the famous LSH scheme MinHash (Broder, 1997) to match tax data and the business register in an elegant and efficient way concerning approximate string matching. The following four paragraphs elaborate on (1) the approximate string matching method for which MinHash is locality-sensitive, (2) creating the MinHash signatures, (3) creating the LSH Forest data structure and (4) the details of our implementation in Python.

The LSH scheme MinHash is locality sensitive for the Jaccard distance on character n-grams, or n-shingles (Leskovec et al., 2014, Chapter 3). A *character n-gram* is defined as a substring of n consecutive characters in a string. As an example, the set of character 3-grams, or trigrams, of the string 'webshop' is the set {'web', 'eb ', 'b s', ' sh', 'sho', 'hop'}. For $n \in \mathbb{N}$, write f_n for the functions mapping a string to its set of character n-grams. The Jaccard distance between two sets A and B is defined as

$$d_J(A, B) = 1 - |A \cap B|/|A \cup B|. \tag{5.14}$$

The n-gram Jaccard distance $d_{J,n}$ between two strings s and t is defined as

$$d_{J,n}(s, t) = 1 - d_J(f_n(s), f_n(t)). \tag{5.15}$$

For MinHash, a string is identified by a binary-valued vector in $\{0, 1\}^{c_n}$, with $c_n = (26 + 10 + 1)^n$ (enumerated in the order of the alphabet (26), digits (10), white space (1)). In fact, a string is stored only as the list of index numbers (according to the n-gram enumeration) of the n-grams it contains. The randomised dimensionality reduction MinHash computes a k-bit min-hash signature for each $f_n(a)$ in the following way. First, randomly choose k hash functions $h_1, \dots, h_k$ from the family of random linear functions of the form $h(x) = (\alpha x + \beta) \bmod p$, with a and b integers and p a fixed, large prime number. Then, randomly choose k

hash functions $g_1, \ldots, g_k$ mapping the values $0, \ldots, p-1$ uniformly at random onto $\{0, 1\}$. The j-th bit of the k-bit min-hash signature of a is then given by $g_j(\min_i\{h_j(v_i)\})$, where v is the list containing the index numbers of the character n-grams of a.

To reduce the number of evaluations of d_{J_n} that are required to match the two data sources, we apply the LSH Forest data structure (Bawa et al., 2005) on the results of MinHash. In short, an *LSH tree* is defined as the logical prefix tree on all k-bit signatures. The LSH forest consists of l LSH trees, each constructed with an independently drawn random sequence of hash functions from the described family of hash functions (MinHash). Now, given the stem s of a company name from the tax data, each of the l LSH trees is updated with an additional leaf node containing (the end point of the path through the LSH tree specified by the k-bit signature of) $s(a)$. The LSH trees are then searched bottom up simultaneously, starting from the new leaf node, until the m most similar items are identified. Consult Bawa et al. (2005) for further algorithmic details.

In our implementation in Python, the function `MinHashLSHForest` from the Python library `datasketch` is used (`https://github.com/ekzhu/datasketch`). We set $n = 3$, i.e., we consider the Jaccard distance of character trigrams. The total number of hash functions is fixed to be 64 and the number of LSH trees was set to the default value $l = 8$. The datasketch implementation then fixes $k = 64/8 = 8$ for the length of the min-hash signatures that are used to build each of the LSH trees. The choice of $k = 8$ is relatively small but works already quite well in our case, as shown in Section 5.4.1. The top $m = 100$ most similar leaf nodes (stems of company names from the business register) from the LSH forest are returned for each stem of company names from the tax data. If the suffix class of a company from the business register is different from that of the company from the tax data, the company from the business register is removed from the list. The resulting lists serve as input for the next part of our data-driven approach for firm-level record linkage.

Step III: combining string distance metrics

We combine the following eight commonly used string distance metrics:

(a) 1, the normalised Levenshtein (or edit) distance;

(b) 2, the Jaro-Winkler divergence (not a metric in the mathematical sense (Winkler, 1990));

(c) 3-5, the Jaccard distance on sets of character 1-, 2- and 3-grams;

(d) 6-8, the cosine distance on term frequency vectors of character 1-, 2- and 3-grams.

The Jaro-Winkler, Jaccard and cosine distances are defined as 1 minus the corresponding string similarity measures and always take values in the interval $[0, 1]$. The Levenshtein distance is normalised to the interval $[0, 1]$, which is achieved by dividing by the maximum length of the two input strings. All metrics are defined and compared by Cohen et al. (2003). For a more recent discussion, see Leskovec et al. (2014, pp. 87-93).

At the end of this step, each company in the tax data is equipped with an eight-dimensional vector containing values in the interval $[0, 1]$. These values can be interpreted as the distance (along different metrics) to the set of EU retail companies in the business register. The values will be used as features in the machine learning algorithms as described in step IV of Section 5.3.1.1.

5.B Finding $\boldsymbol{\lambda}^*$

This appendix describes how the optimal value $\lambda = \lambda^*$ is found. Recall that Van Delden et al. (2016) considered the linear combinations $\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}} = (1-\lambda)\widehat{\boldsymbol{B}}_{\boldsymbol{0}} + \lambda\widehat{\boldsymbol{B}}_{\boldsymbol{1}}$ for $\lambda \in [0, 1]$ of the bias estimators given by Equations (5.8) and (5.9). They proposed to find the optimal value $\lambda = \lambda^*$ that minimises the MSE of the first component $(\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}})_1$ of $\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}}$ as an estimator of the bias $\boldsymbol{B}(\widehat{\boldsymbol{y}})$. This MSE is given by

$$\mathrm{MSE}\left((\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}})_1\right) = \left\{\boldsymbol{B}(\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}})\right\}_1^2 + \left\{V(\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}})\right\}_{11}, \tag{5.16}$$

where $\left\{V(\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}})\right\}_{11}$ denotes the upper left entry in the variance-covariance matrix of the 2-vector $\widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}}$. The following iterative approach is suggested by Van Delden et al. (2016) to find λ^*.

Step 1: Initialise – start with $\lambda_{\mathrm{old}} = 0$.

Step 2: Compute $\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{B}}_{\boldsymbol{0}}$ and $\widehat{\Omega} = \widehat{V}(\widehat{\boldsymbol{y}})$.

Step 3: Compute $\lambda_{\text{new}} = \max\{0, \min\{1, (m_1 - m_3 + m_4)/(m_1 + m_2 - 2m_3 + m_4)\}\}$, where

$$m_1 = \left((\widehat{P}^T - I_2)\widehat{\boldsymbol{B}}\right)_1^2, \tag{5.17}$$

$$m_2 = \left((\widehat{P}^T - I_2)\widehat{\Omega}(\widehat{P}^T - I_2)^T\widehat{Q}^T\widehat{Q}\right)_{11}, \tag{5.18}$$

$$m_3 = \frac{1}{2}\left((\widehat{P}^T - I_2)\widehat{\Omega}(\widehat{P}^T - I_2)^T(\widehat{Q} + \widehat{Q}^T)\right)_{11}, \tag{5.19}$$

$$m_4 = \left((\widehat{P}^T - I_2)\widehat{\Omega}(\widehat{P}^T - I_2)^T\right)_{11}. \tag{5.20}$$

Step 4: If $|\lambda_{\text{new}} - \lambda_{\text{old}}| < 10^{-6}$, stop and return λ_{new}. Otherwise, set

$$\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{B}}_{\boldsymbol{\lambda}_{\text{new}}} = (1 - \lambda_{\text{new}})\widehat{\boldsymbol{B}}_{\mathbf{0}} + \lambda_{\text{new}}\widehat{\boldsymbol{B}}_{\mathbf{1}} = (I_2 + \lambda_{\text{new}}(\widehat{Q} - I_2))\widehat{\boldsymbol{B}}_{\mathbf{0}}. \tag{5.21}$$

Step 5.: Set $\lambda_{\text{old}} := \lambda_{\text{new}}$ and return to step 2.

Details of the derivation of the formulas in step 3 can be found in Appendix A3 in Van Delden et al. (2015). We indicate that the above iterative procedure is performed for each of the years 2014, 2015 and 2016 separately. The same (estimated) matrix $\widehat{P}$ is used for each year. The optimal value of λ might differ across years, as it depends on the annual turnover.

5.C Local behaviour around optimal parameters

This appendix contains additional results on the local behaviour of the mean and standard deviation of F1-scores (obtained from the 5-fold cross-validation) around the optimal parameters for the optimal algorithm. The results for $\widehat{s}_i^{\text{BR}}$ (by the business register) and $\widehat{s}_i^{\text{W}}$ (by websites) are presented separately.

Business Register. Table 5.10 shows that the results for $C \geq 10$ hardly depend on the class-weighting scheme that is chosen. Moreover, different choices of γ and different choices of C, given $C \geq 10$, only minimally affect the mean F1-score over the five folds. The standard deviation is similar in each of these parameter settings as well. Thus, the mean F1-score is stable around the optimal parameter setting.

Table 5.10: Mean (and standard deviation of) F1-scores for RBFSVC predicting $\widehat{s}_i^{\mathrm{BR}}$.[†]

γ	*Results for uniform class weighting*			*Results for balanced class weighting*		
	$C = 1$	$C = 10$	100	$C = 1$	$C = 10$	$C = 100$
0.001	–	0.93 (± 0.05)	0.94 (± 0.02)	–	0.91 (± 0.04)	0.94 (± 0.02)
0.01	0.93 (± 0.05)	0.94 (± 0.02)	0.92 (± 0.03)	0.92 (± 0.05)	0.94 (± 0.02)	0.92 (± 0.03)
0.1	0.94 (± 0.02)	0.93 (± 0.02)	0.96 (± 0.04)	0.94 (± 0.02)	0.93 (± 0.02)	0.96 (± 0.04)
1	0.93 (± 0.03)	0.95 (± 0.04)	**0.97 (± 0.03)**	0.94 (± 0.02)	0.95 (± 0.04)	0.97 (± 0.03)

[†] The optimal parameter setting ($C = 100, \gamma = 1$, uniform) with corresponding F1-score 0.97 is displayed in bold font.

Table 5.11: Mean (and standard deviation of) F1-scores for AB predicting $\widehat{s}_i^{\mathrm{W}}$.[†]

n	*Results for $\lambda = 0.01$, balanced class weighting*			*Results for $\lambda = 0.1$, uniform class weighting*		
	$d = 1$	$d = 2$	$d = 3$	$d = 1$	$d = 2$	$d = 3$
50	0.77 (± 0.06)	0.75 (± 0.13)	0.67 (± 0.15)	0.74 (± 0.14)	0.70 (± 0.17)	0.69 (± 0.12)
100	0.75 (± 0.12)	0.73 (± 0.14)	0.69 (± 0.14)	0.78 (± 0.13)	0.75 (± 0.12)	0.70 (± 0.15)
200	0.75 (± 0.12)	0.71 (± 0.15)	0.67 (± 0.16)	0.77 (± 0.09)	0.73 (± 0.14)	0.67 (± 0.16)
500	0.78 (± 0.10)	0.69 (± 0.15)	0.70 (± 0.15)	0.75 (± 0.10)	0.76 (± 0.13)	0.69 (± 0.15)
	$\lambda = 0.1$, balanced class weighting			*$\lambda = 1$, balanced class weighting*		
	$d = 1$	$d = 2$	$d = 3$	$d = 1$	$d = 2$	$d = 3$
50	0.80 (± 0.09)	0.74 (± 0.15)	0.67 (± 0.12)	0.72 (± 0.08)	0.71 (± 0.06)	0.71 (± 0.14)
100	**0.80 (± 0.11)**	0.75 (± 0.14)	0.65 (± 0.09)	0.71 (± 0.08)	0.71 (± 0.06)	0.70 (± 0.11)
200	0.79 (± 0.10)	0.72 (± 0.13)	0.67 (± 0.09)	0.70 (± 0.07)	0.71 (± 0.10)	0.70 (± 0.11)
500	0.75 (± 0.09)	0.73 (± 0.12)	0.68 (± 0.14)	0.72 (± 0.06)	0.69 (± 0.14)	0.66 (± 0.10)

[†] The optimal parameter setting ($n = 100, d = 1, \lambda = 0.1$, balanced) with corresponding F1-score 0.80 is displayed in bold font.

Websites. Next, we examine the local behaviour of the mean and standard deviation of F1-scores obtained from the 5-fold cross-validation around the optimal parameters for the algorithms AB, GB and RF when predicting $\widehat{s}_i^{\mathrm{W}}$.

Starting with AB, Table 5.11 shows that increasing the maximum tree depth d from the optimal value $d = 1$ negatively impacts the goodness of fit as measured by F1. Moreover, for $\lambda = 0.01$ and $\lambda = 1$, the mean F1-score is substantially lower for $d = 1$ compared with the optimal $\lambda = 0.1$ (all using the balanced class-weighting scheme). Thus, the results by AB are not stable around the optimal maximum tree depth $d = 1$ and not around the optimal learning rate $\lambda = 0.1$. The results are less sensitive to the choice of the class-weighting scheme, for $\lambda = 0.1$.

Studying Table 5.12, we may conclude that the mean F1-scores of GB are not very stable around the optimal parameter setting ($n = 200, d = 1, \lambda = 0.1$). Increasing the maximum depth d from the optimal value $d = 1$ while fixing the

TABLE 5.12: Mean (and standard deviation of) F1-scores for GB predicting $\widehat{s}_i^{W}$.†

n	*Results for* $\lambda = 0.1$ *(d varies)*			*Results for* $d = 1$ *(λ varies)*		
	$d = 1$	$d = 2$	$d = 3$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$
50	0.75 (± 0.12)	0.72 (± 0.14)	0.71 (± 0.15)	0.65 (± 0.05)	0.75 (± 0.12)	0.74 (± 0.10)
100	0.77 (± 0.12)	0.71 (± 0.15)	0.69 (± 0.13)	0.71 (± 0.08)	0.77 (± 0.12)	0.75 (± 0.10)
200	**0.79 (± 0.10)**	0.71 (± 0.15)	0.69 (± 0.13)	0.73 (± 0.12)	**0.79 (± 0.10)**	0.73 (± 0.10)
500	0.74 (± 0.12)	0.72 (± 0.16)	0.66 (± 0.14)	0.76 (± 0.12)	0.74 (± 0.12)	0.73 (± 0.10)

† The optimal parameter setting ($n = 200, d = 1, \lambda = 0.1$) with corresponding F1-score 0.79 is displayed in bold font. Note that the parameter settings for the second and sixth column are identical.

TABLE 5.13: Mean (and standard deviation of) F1-scores for RF predicting $\widehat{s}_i^{W}$.†

n	*Results for uniform class weighting*			*Results for balanced class weighting*		
RF	$d = 1$	$d = 2$	$d = 3$	$d = 1$	$d = 2$	$d = 3$
50	0.75 (± 0.13)	0.72 (± 0.14)	0.74 (± 0.15)	0.74 (± 0.13)	0.76 (± 0.10)	0.71 (± 0.16)
100	0.76 (± 0.13)	0.73 (± 0.14)	0.69 (± 0.17)	0.76 (± 0.13)	0.75 (± 0.13)	0.68 (± 0.15)
200	0.73 (± 0.14)	0.73 (± 0.14)	0.67 (± 0.16)	**0.78 (± 0.10)**	0.71 (± 0.13)	0.71 (± 0.13)
500	0.76 (± 0.14)	0.67 (± 0.14)	0.68 (± 0.13)	0.75 (± 0.10)	0.70 (± 0.15)	0.71 (± 0.12)

† The optimal parameter setting ($n = 200, d = 1$, balanced) with corresponding F1-score 0.78 is displayed in bold font.

optimal learning rate $\lambda = 0.1$, leads to a drop in mean F1-score. The same holds for changing the optimal learning $\lambda = 0.1$ while fixing $d = 1$.

Finally, we present the results of RF on the training data set in Table 5.13. For the balanced class-weighting scheme, the results seem stable as n increases. Moreover, the variance is smaller than in the uniform class-weighting scheme. However, increasing the maximum depth d from the optimal value $d = 1$ leads to a drop in mean F1-score.

CHAPTER 6

SMOOTHED VARIANTS OF THE AUC

6

6.1 Introduction

This thesis so far has focused on official statistics based on *classifiers*. An interesting alternative is to consider official statistics based on *rankers*. Below, we define what a ranker is and why we believe it provides an interesting alternative to classifiers in reducing misclassification bias in statistical learning. This chapter is devoted to a specific theoretical problem concerning model selection of rankers.

6.1.1 Rankers and misclassification bias

In Chapter 2 we have seen that misclassification bias occurs when aggregating the predictions of binary classifiers. In statistical learning, binary classifiers are based on statistical models that for each object produce an estimate of the probability of belonging to the class of interest. Henceforth, we will refer to that probability as the *score*. The score is then cut off at a threshold value c between 0 and 1. Usually, c is set to 0.5 so that cutting off corresponds to rounding off. We claim that rounding off the scores *causes* misclassification bias, even if the model is correctly specified. We prove our claim for LDA (as an example), see the box titled "The LDA Example" below.

The LDA example shows that aggregating the scores instead of counting the classifications prevents misclassification bias *only if* the model assumptions are satisfied. In particular, the LDA example shows that aggregating the scores when dealing with prior probability shift (see Chapter 3) does not prevent misclassification bias. However, Forman (2006) proposed a method to reduce misclassification bias (called *median sweep*) that is based on scores instead of on classifications. The empirical evidence that he provides shows that median sweep outperforms the misclassification estimator $\hat{\alpha}_p$ (see Chapter 2).

Median sweep is a three-step method. In the first step, objects are *ranked* based on the scores produced by the statistical learning method. In the second step, for each threshold $c = b/l$ (for some fixed integer $l \geq 2$ and with $b = 1, \ldots, l-1$), the resulting classifier-based base rate (see also Chapter 2) is estimated by using the misclassification estimator. In the third step, the median of the $l-2$ values from the second step is computed. The median value is final estimate for the base rate.

A statistical learning method that returns (ranked) scores is referred to as a *ranker*. The estimates produces by median sweep are called *ranker-based statistics*. The quality of median sweep's estimate depends on the *ranking performance* of the ranker that is used. Intuitively, a ranker performs well if it attains a score close to 1 to objects belonging to the class of interest and a score close to 0 to objects not belonging to the class of interest. The most common performance metric for rankers is the AUC: the area under the receiver operating characteristic (ROC) curve. The aim is to select the ranker with highest AUC. We next introduce the open problem concerning model selection of rankers.

The LDA Example.

Linear Discriminant Analysis (LDA) is based on the assumption that the vector of features x_i for any data point i, given that the true class label s_i equals k, is a random vector following a multivariate normal distribution with density f_k. The mean μ_k of the distribution differs between classes k, while the covariance matrix $\Sigma = \Sigma_k$ is identical for each class. The covariance matrix Σ and the means μ_k $(k = 1, 2)$, are estimated on the training set. An unlabelled object is classified by computing for each class the probability that it belongs to that class, given the feature values of the object, and then selecting the class to which the highest probability is assigned. For more details, consult Hastie et al. (2009). The proof of the following proposition shows how rounding off such probabilities *causes* misclassification bias.

Proposition 6.1. *Assume that the base rate α^M in the training set is equal to the base rate α is the unlabelled data set. Then, the LDA-estimate $\widehat{\alpha}_L$ is unbiased if and only if there is no class imbalance, i.e., if $\alpha = 0.5$.*

Proof. See Appendix 6.A. An illustration of the proof is given in Fig. 6.1. □

If all modelling assumptions of LDA are satisfied, then aggregating the scores p^M that are predicted by LDA prevents misclassification bias, i.e., $\mathbb{E}[p^M] = \alpha$ (see Lemma 6.4 in Appendix 6.A). However, if prior probability shift occurs, i.e., $\alpha^M \neq \alpha$, then aggregating scores also results in misclassification bias, see Proposition 6.2 below.

Proposition 6.2. *If $\alpha^M \neq \alpha$, then the expectation $\mathbb{E}[p^M]$ does not equal α, in general.*

Proof. See Appendix 6.A. □

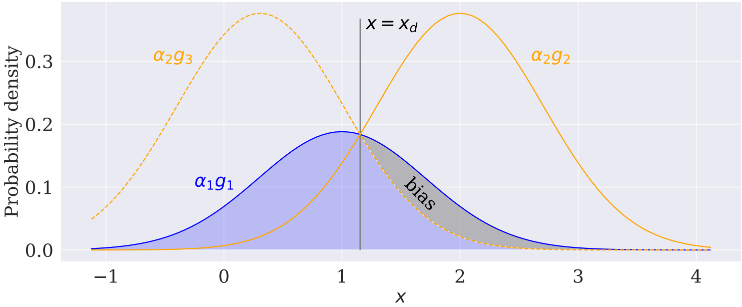


Fig. 6.1: Illustration of how rounding off scores causes misclassification bias in The LDA Example. The grey area is the misclassification bias that occurs when estimating the blue area with LDA as classifier. This figure is the geometric interpretation of the proof of Proposition 6.1, see Appendix 6.A.

6.1.2 Model selection of rankers

Performance metrics for classification algorithms (such as accuracy or the F1-score) often require categorical data as input. Therefore, they cannot evaluate the quality of the score distribution produced by a ranker. The exceptions are performance metrics that are based on, for example, the ROC curve. Such metrics integrate classification performance over all possible threshold values. The AUC is indeed the most commonly used metric, but many alternatives exist (Majnik & Bosnić, 2013). A particularly appealing alternative is the area under the precision-recall curve. It leads to different optimisation results than AUC, so the two are not equivalent (Davis & Goadrich, 2006). Moreover, the area under the precision-recall curve is claimed to be more suitable than AUC for class-imbalanced data sets (Sofaer, Hoeting & Jarnevich, 2019).

In this chapter we will investigate the AUC (and not the alternatives) because of its close relation to the Wilcoxon-Mann-Whitney statistic (Mann & Whitney, 1947; Wilcoxon, 1945): the classical estimator of the AUC is identical to the ($[0, 1]$-normalised) Wilcoxon-Mann-Whitney statistic (Bamber, 1975). We will refer to that estimator as the *standard AUC estimator*. Many theoretical properties of the Wilcoxon-Mann-Whitney test statistic are known. As we are interested in theoretical properties of model selectors, we therefore prefer to use that AUC.

A closer look into the Wilcoxon-Mann-Whitney statistic shows that it is computed by summing the outcome of an indicator function, which is not differentiable. To improve optimisation, Yan et al. (2003) proposed to approximate the indicator function with a smooth function. The resulting variant of the standard AUC will be referred to as the *softAUC estimator* (see Vanderlooy & Hüllermeier, 2008). The open problem is the following: is the softAUC estimator a better estimator of the true AUC of rankers and, ultimately, does a sotftAUC estimator result in the selection of better rankers?

The starting point of our research will be the paper by Vanderlooy and Hüllermeier (2008). They claim that the standard AUC most often selects better rankers, supported by a theoretical analysis combined with empirical evidence. However, we believe that their theoretical analysis should be improved and that their empirical evidence is insufficient to support their claim. In fact, we conjecture that the opposite of their claim holds for any base rate and any set of rankers to select from. We postulate our conjecture in two parts. The first part of our conjecture states that a specific variant of the standard AUC estimator , under some regularity conditions, can be tuned to have a smaller MSE than the standard AUC estimator when estimating the AUC of rankers. The crux is to let the parameter β of the softAUC estimator be dependent on the underlying distribution of the data (see Fig. 6.2). We provide a complete proof of the first part of our conjecture for data sets containing only a single observation for each class. Moreover, we provide some first suggestions for how to generalise the proof to larger data sets. The second part of our conjecture states that the softAUC estimator can also be tuned to be better model selector of rankers than the standard AUC estimator. Again, we provide a partial proof only, including suggestions for how to complete it.

We stress that this chapter is a theoretical contribution. In practice, in particular for larger data sets, the MSE of the standard AUC estimator will be (very) close to 0, see also the upper bounds derived by Birnbaum and Klose (1957). For smaller data sets, the MSE of the standard AUC estimator could be improved more substantially by employing softAUC estimators. However, the data set might be too small to tune the additional parameter β, so other solutions are recommended (Airola, Pahikkala, Waegeman, De Baets & Salakoski, 2009). Still, our theoretical re-evaluation of AUC estimators could indicate that the use of softAUC estimators is justified in applications. At the least, our contribution motivates a further theoretical analysis of softAUC estimators.

Moreover, we believe that our theoretical results are new, as most existing

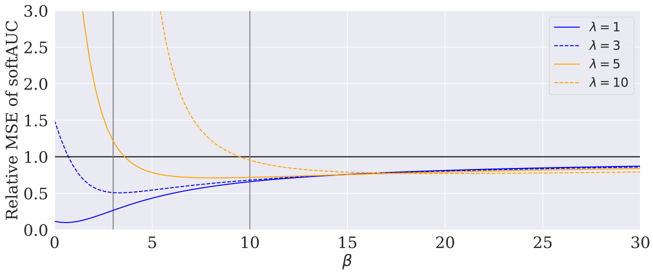


FIG. 6.2: Duplication of the simulation study by Vanderlooy and Hüllermeier (2008). They considered $1 - X$ and Y to be the exponential distributions with rate parameter λ (in their notation: α), restricted to the interval $[0, 1]$. The plot shows the MSE of the softAUC estimator $\widehat{\eta}_\beta$ relative to that of the standard AUC estimator $\widehat{\eta}$ (assuming a data set containing a single observation for each class) for different values of α. The horizontal black line is drawn at $y = 1$; this is where the MSE of $\widehat{\eta}_\beta$ is equal to that of $\widehat{\eta}$. In contrast to fixing $\beta = 3$ or $\beta = 10$ (vertical gray lines) a priori, we illustrate how the points of intersection increase with α, and that the MSE of $\widehat{\eta}_\beta$ is smaller than that of $\widehat{\eta}$, for β sufficiently large.

theoretical results are focused on the standard AUC estimator (i.e., the Wilcoxon-Mann-Whitney statistic) and not on the softAUC estimator. Existing results include the proof that the standard AUC estimator (for data sets containing a single observation for each class) attains the Cramér-Rao lower bound and hence is the uniformly minimum-variance unbiased (UMVU) estimator of the AUC of rankers (Lenstra, 2005, Section 5). However, we will show that the bias that is introduced by smoothing decreases the variance sufficiently to reduce the MSE. Furthermore, we present results for data sets of finite size, whereas existing results are mostly asymptotic by viewing the Wilcoxon-Mann-Whitney statistic as a two-sample U-statistic (Van der Vaart, 1998, Chapter 12).

The remainder of this chapter is organised as follows. In Section 6.2, we provide the formal mathematical definitions of the standard AUC and softAUC estimators and we postulate our two-part conjecture. In Section 6.3, we prove the first part of our conjecture for data sets containing a single observation for each class. In Section 6.4, we propose how to generalise our approach to larger data sets and provide a full proof of the first part of the conjecture for a uniform score distribution and a specific scaling function. In Section 6.5, we discuss the second part of our conjecture and provide a partial proof. Finally, in Section 6.6, we present our conclusions and recommend directions for future research.

6.2 Preliminaries and our conjecture

Assume that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent draws from two different score distributions, corresponding to the positive and negative labels, respectively. We write F_X and F_Y for the respective cumulative distributions functions. Consider the Wilcoxon-Mann-Whitney statistic (Mann & Whitney, 1947; Wilcoxon, 1945)

$$\widehat{\eta}(X_1, \ldots, X_n, Y_1, \ldots, Y_m) := \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} w_*(X_i - Y_j), \tag{6.1}$$

where $w_*(t)$ equals 1 for $t > 0$, equals 0 for $t < 0$ and equals 1/2 for $t = 0$. The expectation of $\widehat{\eta}$ is denoted by γ and can be expressed by the double integral

$$\gamma := \int_0^1 \int_0^1 w_*(x - y) f_X(x) f_Y(y)\, dx\, dy. \tag{6.2}$$

It easily follows that γ is the true AUC of the ranker corresponding to F_X and F_Y, by observing that the points on the ROC curve can be parameterised by $(1 - F_Y(t), 1 - F_X(t))$, for $t \in [0, 1]$. As noted in Section 6.1.1, this observation has been made by Bamber (1975) for the first time. Hence, $\widehat{\eta}$ is an unbiased estimator of the AUC. The estimator in Equation (6.1) will be referred to as the *standard AUC estimator*.

We compare the standard AUC estimator with *softAUC estimators*. They are parameterised by a scalar $\beta \geq 0$ and are defined as

$$\widehat{\eta}_\beta(X_1, \ldots, X_n, Y_1, \ldots, Y_m) := \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} w_\beta(X_i - Y_j), \tag{6.3}$$

in which $w_\beta(t) = f(\beta t)$ with f any increasing function mapping *onto* the open interval $(0, 1)$ that is point symmetric at $t = 0$. We will refer to such functions as *point-symmetric scaling functions*. Yan et al. (2003) first proposed such a softAUC estimator. They used the logistic function $f(t) = 1/(1 + \exp(-t))$ as point-symmetric scaling function, but other functions can be used as well. In fact, the set of point-symmetric scaling functions contains (shifted and scaled) sigmoid functions. We remark that all point-symmetric scaling functions f satisfy the identities $f(0) = 1/2$, $\lim_{t\to\infty} f(t) = 1$, and $\lim_{t\to-\infty} f(t) = 0$.

The expectation of $\widehat{\eta}_\beta$ is denoted by γ_β. Observe that $w_\beta(t) \longrightarrow w_*(t)$ as $\beta \longrightarrow \infty$, for all $t \in [-1, 1]$, and hence $\gamma_\beta \longrightarrow \gamma$.

With the notation partially adopted from Van Dantzig (1951), the MSE of the standard AUC estimator can be expressed as

$$M_*(n,m) := \mathrm{MSE}(\widehat{\eta}(X_1,\dots,X_n,Y_1,\dots,Y_m)) = \frac{1}{nm}\left(\tau_*^2 + (n-1)\phi_*^2 + (m-1)\psi_*^2\right), \tag{6.4}$$

in which

$$\tau_*^2 = \mathbb{E}[(w_*(X_1 - Y_1) - \gamma)^2], \tag{6.5}$$
$$\phi_*^2 = \mathbb{E}[(w_*(X_1 - Y_1) - \gamma)(w_*(X_2 - Y_1) - \gamma)], \tag{6.6}$$
$$\psi_*^2 = \mathbb{E}[(w_*(X_1 - Y_1) - \gamma)(w_*(X_1 - Y_2) - \gamma)]. \tag{6.7}$$

Similarly, the MSE of the softAUC estimator can be expressed as

$$\begin{aligned} M_\beta(n,m) &:= \mathrm{MSE}(\widehat{\eta}_\beta(X_1,\dots,X_n,Y_1,\dots,Y_m)) \\ &= \frac{1}{nm}\left(\tau_\beta^2 + (n-1)\phi_\beta^2 + (m-1)\psi_\beta^2 + (n-1)(m-1)(\gamma-\gamma_\beta)^2\right), \end{aligned} \tag{6.8}$$

in which

$$\tau_\beta^2 = \mathbb{E}[(w_\beta(X_1 - Y_1) - \gamma)^2], \tag{6.9}$$
$$\phi_\beta^2 = \mathbb{E}[(w_\beta(X_1 - Y_1) - \gamma)(w_\beta(X_2 - Y_1) - \gamma)], \tag{6.10}$$
$$\psi_\beta^2 = \mathbb{E}[(w_\beta(X_1 - Y_1) - \gamma)(w_\beta(X_1 - Y_2) - \gamma)]. \tag{6.11}$$

We conjecture that the difference $M_*(n,m) - M_\beta(n,m)$ is positive, for β sufficiently large.

Conjecture 6.1. *For any pair of score distributions F_X and F_Y that are absolutely continuous with respect to the Lebesgue measure on $[0,1]$ and that have overlapping support, and for any increasing point-symmetric (at $t = 0$) function f that maps onto $(0,1)$ and is square-integrable on $(-\infty, 0]$, and for any $n, m \geq 1$, there exist constants $B_{n,m} > 0$ and $C_{n,m} > 0$ such that*

$$M_*(n,m) - M_\beta(n,m) > \frac{C_{n,m}}{\beta} \tag{6.12}$$

for all $\beta > B_{n,m}$.

In Section 6.3, we will prove that positive constants A and B_τ exist such that $\tau_*^2 - \tau_\beta^2 > A/\beta$ for all $\beta > B_\tau$ (in Theorem 6.1). Moreover, we will prove that a positive constant D exists such that $(\gamma - \gamma_\beta)^2 < D/\beta^2$ for all $\beta > 0$ (in Theorem 6.2). Subsequently, we will briefly discuss an example to show that the square integrability of f on $(-\infty, 0]$ is a necessary condition. By combining Theorem 6.1 and Theorem 6.2 we will prove Conjecture 6.1 for $n = m = 1$. In Section 6.4, we suggest how to generalise to $n, m > 1$ and we prove the conjecture for $n, m > 1$ and a specific choice of F_X, F_Y and f.

We postulate that Conjecture 6.1 implies the following result, which is discussed in Section 6.5.

Conjecture 6.2. *If Conjecture 6.1 is true, then η_β is a better model selector than η, for β sufficiently large. More precisely, let S^A and S^B represent random vectors of $n + m$ scores of two models A and B (trained on the same data set) with true AUCs satisfying $\gamma_A > \gamma_B$. Then, we conjecture that there exists some constant $A_{n,m} > 0$ such that*

$$\mathbb{P}\left(\eta_\beta(S^A) - \eta_\beta(S^B) > 0\right) > \mathbb{P}\left(\eta(S^A) - \eta(S^B) > 0\right), \tag{6.13}$$

for all $\beta > A_{n,m}$.

6.3 General theoretical results

If X and Y are independent and have CDFs F_X and F_Y, respectively, then denote the CDF of $T := X - Y$ by G. The distribution of T is referred to as the *score margin distribution* (Vanderlooy & Hüllermeier, 2008). We assume that F_X, F_Y and (hence) G are absolutely continuous with respect to the Lebesgue measure on the Euclidean subspace $[-1, 1]$ (i.e., differentiable real-valued functions) with corresponding probability density functions (PDFs) f_X, f_Y and g. It follows that

$$\gamma = \int_{-1}^{1} w_*(t) g(t)\, dt, \quad \text{and} \quad \gamma_\beta = \int_{-1}^{1} w_\beta(t) g(t)\, dt. \tag{6.14}$$

Below, we will show that $\tau_*^2 - \tau_\beta^2 > A/\beta$ for some $A > 0$ and β sufficiently large (Subsection 6.3.1) and that $(\gamma - \gamma_\beta)^2 < D/\beta^2$ for some $D > 0$ and β sufficiently large (Subsection 6.3.2). For the reader's convenience we reiterate (see Section 6.1) that the differences $\phi_*^2 - \phi_\beta^2$ and $\psi_*^2 - \psi_\beta^2$ are discussed in Section 6.4.

6.3.1 Difference of the first variance terms

The proof of Theorem 6.1 is the main result of the chapter. It shows that the difference $\tau_*^2 - \tau_\beta^2$ converges to 0 from above as $\beta \longrightarrow \infty$, and that the convergence rate is at most $1/\beta$. Compared with Vanderlooy and Hüllermeier (2008), we add the assumption that $g(0) > 0$, which we believe is a reasonable assumption. If the assumption does not hold, we either have $\gamma = 1$, or, if $\gamma < 1$, we need the sets A_X and A_Y to which f_X and f_Y assign positive probability to have *empty* intersection, but still contain points $x \in A_X$ and $y \in A_Y$ with $x < y$. Both cases seem unlikely to occur in practice. In contrast to Vanderlooy and Hüllermeier (2008), we do *not* need to assume that $g(-t) \leq g(t)$ for all $t \in [0, 1]$. In fact, we do not need g to be continuous at any other point than at $t = 0$, as long as g is bounded on $[-1, 1]$. In typical applications, g will be continuous on $[-1, 1]$, implying that it is also bounded on $[-1, 1]$.

Theorem 6.1. *If g is continuous at $t = 0$ with $g(0) > 0$, then for any point-symmetric scaling function f, there exist constants $B_\tau \geq 0$ and $A > 0$ such that $\tau_*^2 - \tau_\beta^2 > A/\beta$ for all $\beta > B_\tau$.*

Proof. If a ranker has true AUC γ, then the ranker that is obtained by applying the transformation $x \mapsto 1 - x$ to the scores has true AUC $1 - \gamma$. Therefore, we may assume (without loss of generality) that the true AUC γ is at least $1/2$. We distinguish two cases.

The first case. Assume that $1/2 \leq \gamma \leq 3/4$. Let $B_\tau = 0$ and $\beta > B_\tau$. For $-1 \leq t < 0$, it holds that $0 < w_\beta(t) \leq 1/2$. Hence, as $\gamma \geq 1/2$, $|w_\beta(t) - \gamma| < \gamma = |w_*(t) - \gamma|$. In addition, for $0 < t \leq 1$, it holds that $1/2 \leq w_\beta \leq 1$. It implies, as $\gamma \leq 3/4$, that for $t > 0$

$$|w_\beta(t) - \gamma| \leq \max\{|1/2 - \gamma|, |1 - \gamma|\} = |1 - \gamma| = |w_*(t) - \gamma|. \tag{6.15}$$

Hence, $|w_\beta(t) - \gamma| \leq |w_*(t) - \gamma|$ and thus $(w_\beta(t) - \gamma)^2 \leq (w_*(t) - \gamma)^2$ for all $-1 \leq t \leq 1$, with strict inequality for all $t \neq 0$ and equality for $t = 0$, as $w_\beta(0) = 1/2 = w_*(0)$. We obtain a lower bound for the difference of the first variance terms, given by

$$\tau_*^2 - \tau_\beta^2 = \int_{-1}^{1} (w_*(t) - \gamma)^2 - (w_\beta(t) - \gamma)^2 g(t)\, dt \geq \int_U (w_*(t) - \gamma)^2 - (w_\beta(t) - \gamma)^2 g(t)\, dt, \tag{6.16}$$

for any subset $U \subset [-1,1]$. We choose a suitable U as follows. As g is continuous at $t = 0$ with $g(0) > 0$, there is a $\delta > 0$ such that $g(t) > g(0)/2 > 0$ for all t in the open interval $(-\delta, \delta)$. Then take $U = U_\beta = [0, \delta/\beta]$. To compute the integral over U, define the primitive functions

$$F_1(t) := \int_0^t f(s)\, ds, \quad \text{and} \quad F_2(t) := \int_0^t (f(s))^2\, ds. \tag{6.17}$$

Verify, by substituting $r = \beta s$, that

$$\int_{U_\beta} w_\beta(s)\, ds = \int_0^\delta \frac{1}{\beta} w_\beta(r/\beta)\, dr = \frac{1}{\beta}\int_0^\delta f(r)\, dr = \frac{1}{\beta}F_1(\delta). \tag{6.18}$$

Similarly, $\int_{U_\beta}(w_\beta(s))^2\, ds = F_2(\delta)/\beta$. The lower bound on $\tau_*^2 - \tau_\beta^2$ then becomes

$$\begin{aligned}
\tau_*^2 - \tau_\beta^2 &> \frac{g(0)}{2}\int_{U_\beta} (w_*(t) - \gamma)^2 - (w_\beta(t) - \gamma)^2\, dt \\
&= \frac{g(0)}{2\beta}\left[\delta - 2\gamma\delta + \gamma^2\delta - F_2(\delta) + 2\gamma F_1(\delta) - \gamma^2\delta\right] \\
&= \frac{g(0)}{2\beta}\left[\delta - F_2(\delta) + 2\gamma(F_1(\delta) - \delta)\right] \\
&\geq \frac{g(0)}{2\beta}\left[F_1(\delta) - F_2(\delta)\right],
\end{aligned} \tag{6.19}$$

where the last equality follows from $\gamma \geq 1/2$. Note that $F_1(\delta) - F_2(\delta) > 0$ because $0 \leq f \leq 1$, hence $f^2 \leq f$, and $\delta > 0$. We take $B_\tau = 0$ and $A = F_1(\delta) - F_2(\delta)$, which is indeed a constant that does not depend on β, to complete the proof for the first case.

The second case. Assume that $3/4 < \gamma \leq 1$. The continuity of g at $t = 0$ with $g(0) > 0$ guarantees that $\gamma < 1$. We use a second ϵ-δ-argument to complete the proof. Let ϵ be equal

$$\epsilon = \frac{1}{2}\frac{F_1(t_\gamma) - F_2(t_\gamma)}{(2\gamma - 1)(t_\gamma - F_1(t_\gamma))} \cdot g(0). \tag{6.20}$$

In the expression for ϵ, t_γ is defined as the positive solution to the equation

$(f(t) - \gamma)^2 = (w_* - \gamma)^2$, which exists (and is unique) if and only if $3/4 < \gamma < 1$. The functions F_1 and F_2 are the primitive function of f and f^2, respectively, as defined before. As f maps into $(0, 1)$ and because t_γ is positive, it follows that $t_\gamma - F_1(t_\gamma) > 0$ as well as $F_1(t_\gamma) - F_2(t_\gamma) > 0$. The assumption $g(0) > 0$ implies that $\epsilon > 0$. As g is continuous at $t = 0$, there exists a $\delta > 0$ such that $|g(t) - g(0)| < \epsilon$ for all t in the open interval $(-\delta, \delta)$. Next, we take $B_\tau > 0$ as

$$B_\tau = \frac{1}{\delta} t_\gamma. \tag{6.21}$$

Our task is now to show that $\tau_*^2 - \tau_\beta^2 > A/\beta$ for some $A > 0$ and any $\beta > B_\tau$.

So, let $\beta > B_\tau$. Analogously to the first case, it holds that $(w_\beta(t) - \gamma)^2 < (w_*(t) - \gamma)^2$ for $-1 \le t < 0$. Solving $(w_\beta(t) - \gamma)^2 = (w_*(t) - \gamma)^2$ for $t > 0$ yields the unique solution

$$t' = \frac{1}{\beta} t_\gamma. \tag{6.22}$$

Note that $t' < \delta$. Moreover, it holds that $(w_\beta(t) - \gamma)^2 < (w_*(t) - \gamma)^2$ for all $t \in (t', 1]$, by definition of t'. It leads to the lower bound

$$\begin{aligned} \tau_*^2 - \tau_\beta^2 &= \int_{-1}^{1} \left[(w_*(t) - \gamma)^2 - (w_\beta(t) - \gamma)^2 \right] g(t)\, dt \\ &> \int_{-t'}^{t'} \left[(w_*(t) - \gamma)^2 - (w_\beta(t) - \gamma)^2 \right] g(t)\, dt. \end{aligned} \tag{6.23}$$

Split the latter integral into the part left of $t = 0$ and the part right of $t = 0$, and denote them by I_L and I_R, respectively. Note that I_L is positive and I_R is negative. The aim is to derive bounds for I_L and I_R. To that end, first verify, by substituting $r = \beta s$ as before, that

$$\int_0^{t'} w_\beta(s)\, ds = \int_0^{t_\gamma} \frac{1}{\beta} w_\beta(r/\beta)\, dr = \frac{1}{\beta} \int_0^{t_\gamma} f(r)\, dr = \frac{1}{\beta} F_1(t_\gamma). \tag{6.24}$$

Similarly, $\int_0^{t'} (w_\beta(s))^2\, ds = F_2(t_\gamma)/\beta$. Furthermore, note that $0 < g(t) < g(0) + \epsilon$ for all $t \in [0, t']$. An upper bound on the positive number $-I_R$ is then given by

$$
\begin{aligned}
-I_R &= \int_0^{t'} \left[(w_\beta(t))^2 - 2\gamma w_\beta(t) + \gamma^2 - \gamma^2 + 2\gamma - 1\right] g(t)\, dt \qquad (6.25)\\
&< (g(0)+\epsilon)\left[\frac{1}{\beta}F_2(t_\gamma) - \frac{2\gamma}{\beta}F_1(t_\gamma) + \frac{(2\gamma-1)}{\beta}t_\gamma\right] \qquad (6.26)\\
&= \frac{g(0)+\epsilon}{\beta}\Big((2\gamma-1)(t_\gamma - F_1(t_\gamma)) - (F_1(t_\gamma) - F_2(t_\gamma))\Big). \qquad (6.27)
\end{aligned}
$$

Next, the point-symmetry of f at $t = 0$ can be used to show that

$$
\begin{aligned}
\int_{-t'}^{0} w_\beta(s)\, ds = \frac{1}{\beta}\int_{-t_\gamma}^{0} f(r)\, dr &= \frac{1}{\beta}\int_0^{t_\gamma} f(-r)\, dr\\
&= \frac{1}{\beta}\int_0^{t_\gamma}(1-f(r))\, dr = \frac{1}{\beta}\left(t_\gamma - F_1(t_\gamma)\right). \qquad (6.28)
\end{aligned}
$$

Similarly, it follows that

$$
\int_{-t'}^{0}(w_\beta(s))^2\, ds = \frac{1}{\beta}\int_0^{t_\gamma}\left(1-f(r)\right)^2\, dr = \frac{1}{\beta}\left(t_\gamma - 2F_1(t_\gamma) + F_2(t_\gamma)\right). \qquad (6.29)
$$

Then, one can derive that

$$
\begin{aligned}
I_L &:= \int_{-t'}^{0}\left[(w_*(t)-\gamma)^2 - (w_\beta(t)-\gamma)^2\right] g(t)\, dt\\
&> \frac{g(0)-\epsilon}{\beta}\int_{-t'}^{0}\left[-(w_\beta(s))^2 + 2\gamma w_\beta(s)\right]\, ds\\
&= \frac{g(0)-\epsilon}{\beta}\Big(-t_\gamma + 2F_1(t_\gamma) - F_2(t_\gamma) + 2\gamma t_\gamma - 2\gamma F_2(t_\gamma)\Big)\\
&= \frac{g(0)-\epsilon}{\beta}\Big((2\gamma-1)(t_\gamma - F_1(t_\gamma)) + (F_1(t_\gamma) - F_2(t_\gamma))\Big). \qquad (6.30)
\end{aligned}
$$

Introducing the notation $a = (2\gamma - 1)(t_\gamma - F_1(t_\gamma))$ and $b = F_1(t_\gamma) - F_2(t_\gamma)$, observe that $\epsilon = b/2a \cdot g(0)$ and hence

$$\begin{aligned} I_L + I_R &> \frac{1}{\beta}\big((g(0) - \epsilon)(a+b) - (g(0)+\epsilon)(a-b)\big) \\ &= \frac{g(0)}{\beta}\left(\frac{2a-b}{2a} \cdot (a+b) - \frac{2a+b}{2a} \cdot (a-b)\right) \\ &= \frac{g(0)}{\beta} \cdot \frac{a(a+b) - a(a-b)}{2a} = \frac{b \cdot g(0)}{\beta}. \end{aligned} \tag{6.31}$$

We take $A = b \cdot g(0) = (F_1(t_\gamma) - F_2(t_\gamma)) \cdot g(0) > 0$. Hence, for this choice of $A > 0$ and for any $\beta > B_\tau$ we may conclude that $\tau_*^2 - \tau_\beta^2 > A/\beta$. It completes the proof for the second case and thus concludes the proof of the theorem. □

We point out that the entire proof can be used to extend the result to the difference of the first term in the mean *absolute* error of the two estimators under consideration. The only modification is choosing $2\epsilon = (1-\gamma)t_\gamma/(\gamma t_\gamma - F_1(t_\gamma)) \cdot g(0)$ instead. Moreover, the convergence rate of at most $1/\beta$ that we obtained in Theorem 6.1 above has been essential in postulating Conjecture 6.1, as will become clear in the next subsection.

6.3.2 Convergence rate of squared bias

The result below shows that the squared bias term $(\gamma - \gamma_\beta)^2$ in the MSE of the softAUC estimator, see expression (6.8), converges to 0 as $\beta \to \infty$ with a convergence rate of at least $1/\beta^2$.

Theorem 6.2. *Assume that g is bounded and that f is a point-symmetric scaling function, square integrable on $(-\infty, 0]$. Then, $(\gamma - \gamma_\beta)^2 < D/\beta^2$ for some constant $D > 0$ and all $\beta > 0$.*

Proof. Let R be an upper bound on g and let $c = \int_{-\infty}^{0} (f(t))^2 \, dt$, which is finite by assumption. An upper bound on the squared difference $(\gamma - \gamma_\beta)^2$ is then given by

$$\begin{aligned}(\gamma - \gamma_\beta)^2 &= \left(\int_{-1}^{1} (w_*(t) - w_\beta(t)) g(t) \, dt \right)^2 \\ &\leq \int_{-1}^{1} (w_*(t) - w_\beta(t))^2 \cdot (g(t))^2 \, dt \\ &\leq 2R^2 \int_{-1}^{0} (f(\beta t))^2 \, dt = \frac{2R^2}{\beta^2} \int_{-\beta}^{0} (f(t))^2 \, dt < \frac{2cR^2}{\beta^2}. \end{aligned} \tag{6.32}$$

We take $D = 2cR^2$. This completes the proof. □

Combining Theorem 6.1 with Theorem 6.2 proves Conjecture 6.1 for the case of $n = m = 1$ by

$$M_*(1,1) - M_\beta(1,1) = \tau_*^2 - \tau_\beta^2 - (\gamma - \gamma_\beta)^2 > \frac{A}{\beta} - \frac{D}{\beta^2} > \frac{A/2}{\beta}, \tag{6.33}$$

for all $\beta > \max\{2D/A, B_\tau\}$.

Finally, we take a closer look at the condition of square integrability in Theorem 6.2 and make two observations. First, we observe that the condition $\int_{-\infty}^{0} (f(t))^2 \, dt < \infty$ is satisfied by the logistic function $f(t) = 1/(1 + \exp(-t))$, as the improper integral exists and is equal to $\log(2) - 1/2$. Our second observation is that square integrability is also a necessary condition for Conjecture 6.1 to hold for larger values of n and m. For a counterexample, we take $f(t) = 1 - 1/\sqrt{4+t}$ for $t \geq 0$ and $f(t) = 1/\sqrt{4-t}$ for $t \leq 0$. This function is not square integrable. We take the uniform distribution on $[-1/3, 1]$ as our distribution G. It can be shown that

$$\gamma - \gamma_\beta = \frac{3}{2} \frac{1}{\sqrt{\beta}} \left(\sqrt{1 + 4/\beta} - \sqrt{1/3 + 4/\beta} \right) > \frac{3}{8} \frac{1}{\sqrt{\beta}}, \tag{6.34}$$

where the inequality holds for all $\beta > 2304/649 \approx 3.55$. Hence, $(\gamma - \gamma_\beta)^2 > D/\beta$, which implies that the difference in MSE (see inequality (6.12)) eventually becomes negative, for β, n, and m sufficiently large. Indeed, this counterexample shows that square integrability is a necessary condition for Conjecture 6.1 to hold for larger values of n and m.

6.4 Analytic proof for a simple score function

In this section we will investigate the sign and convergence rate of both $\phi_*^2 - \phi_\beta^2$ and $\psi_*^2 - \psi_\beta^2$. In Subsection 6.4.1, we prove that if X and $1 - Y$ have the same distribution, then $\phi_*^2 - \phi_\beta^2 = \psi_*^2 - \psi_\beta^2$, hence in that case it suffices to investigate $\psi_*^2 - \psi_\beta^2$ only. In Subsection 6.4.2, for a simple score distribution and an algebraic scaling function, we prove analytically that $\psi_*^2 - \psi_\beta^2 > -C/\beta^2$ for some constant $C > 0$ and any $\beta > 0$. It is sufficient to prove Conjecture 6.1 for the example.

6.4.1 A useful symmetry argument

Lemma 6.1 formulates a symmetry argument which is useful in reducing the computations in Subsection 6.4.2.

Lemma 6.1. *If X and $1 - Y$ have the same distribution, then $\phi_*^2 = \psi_*^2$ and $\phi_\beta^2 = \psi_\beta^2$.*

Proof. The assumption that X and $1 - Y$ have the same distribution implies that $f_X(x) = f_Y(1 - x)$. The result then follows directly from substituting $s = 1 - x$ and $t = 1 - y$ in the expression for ϕ_*^2 or ϕ_β^2. For completeness,

$$\begin{aligned}
\phi_*^2 &= \int_0^1 \left(\int_0^1 (w_*(x - y) - \gamma) f_X(x)\, dx \right)^2 f_Y(y)\, dy \\
&= \int_0^1 \left(\int_0^1 (w_*(x - y) - \gamma) f_Y(1 - x)\, dx \right)^2 f_X(1 - y)\, dy \\
&= -\int_1^0 \left(-\int_1^0 (w_*(1 - s - (1 - t)) - \gamma) f_Y(s)\, ds \right)^2 f_X(t)\, dt \\
&= \int_0^1 \left(\int_0^1 (w_*(t - s) - \gamma) f_Y(s)\, ds \right)^2 f_X(t)\, dt = \psi_*^2. \qquad (6.35)
\end{aligned}$$

The proof of $\phi_\beta^2 = \psi_\beta^2$ is identical. □

In the next subsection, we prove Theorem 6.3, which states that $\psi_*^2 - \psi_\beta^2 > -C/\beta^2$ for a uniform score distribution.

6.4.2 Complete proof for uniform score distribution

So far, we have not yet succeeded in deriving general bounds for the convergence rate of $\psi_*^2 - \psi_\beta^2$ that can be used to prove Conjecture 6.1. Here, we study that convergence rate for a specific example, see the box titled "The Uniform Example".

The Uniform Example.

To investigate the convergence rate of $\psi_*^2 - \psi_\beta^2$ we consider a simple example, namely a uniform score distribution in combination with an algebraic point-symmetric scaling function. The specific example that we consider is as follows.

Assume that X and Y are independent random variables that follow a uniform distribution on $[1/3, 1]$ and $[0, 2/3]$, respectively. Let f be the following point-symmetric scaling (and sigmoid) function:

$$f(t) = \frac{1}{2}\frac{1}{\sqrt{1+t^2}} + \frac{1}{2}. \tag{6.36}$$

We are able to prove Conjecture 6.1 for this specific example.

We begin by computing γ and γ_β.

Lemma 6.2. *In The Uniform Example, we have* $\gamma = 7/8$ *and*

$$\gamma_\beta = \frac{1}{2} + \frac{9}{16}\sqrt{1+1/\beta^2} - \frac{3}{16}\sqrt{1+9/\beta^2} + \frac{9}{16\beta^2}\log(3) - \frac{9}{16\beta^2}\log\left(\frac{1+\sqrt{1+9/\beta^2}}{1+\sqrt{1+1/\beta^2}}\right). \tag{6.37}$$

Observe that the limit of γ_β *for* $\beta \to \infty$ *equals* $7/8$ $(= \gamma)$*, as it should.*

Proof. The true AUC γ can be computed directly, yielding

$$\begin{aligned} \gamma &= \int_0^1 \int_0^1 w_*(x-y)\, dF_Y(y) dF_X(x) \\ &= \left(\frac{3}{2}\right)^2 \cdot \int_{1/3}^{2/3}\int_0^x 1\, dy\, dx + \left(\frac{3}{2}\right)^2 \cdot \int_{2/3}^{1}\int_0^{2/3} 1\, dy\, dx = \frac{3}{8} + \frac{1}{2} = \frac{7}{8}. \end{aligned} \tag{6.38}$$

Next, let $\beta > 0$.The expected value γ_β of the softAUC can be expressed as

$$\begin{aligned}\gamma_\beta &= \int_0^1 \int_0^1 w_\beta(x-y)\, dF_Y(y)\, dF_X(x) = \frac{9}{4}\int_{1/3}^1 \int_0^{2/3} \frac{1}{2}\frac{\beta(x-y)}{\sqrt{1+(\beta(x-y))^2}} + \frac{1}{2}\, dy\, dx \\ &= \frac{9}{8}\int_{1/3}^1 \int_0^{2/3} \frac{\beta(x-y)}{\sqrt{1+(\beta(x-y))^2}}\, dy\, dx + \frac{1}{2}. \end{aligned} \tag{6.39}$$

To compute the inner integral, note that, for any real number x, a primitive function of $w_\beta(x-y)$ with respect to y is given by

$$y \mapsto -\frac{1}{\beta}\sqrt{1+(\beta(x-y))^2}. \tag{6.40}$$

Hence,

$$\int_0^{2/3} \frac{\beta(x-y)}{\sqrt{1+(\beta(x-y))^2}}\, dy = \frac{1}{\beta}\left(\sqrt{1+(\beta x)^2} - \sqrt{1+(\beta(x-\tfrac{2}{3}))^2}\right). \tag{6.41}$$

Then, for any real number a, observe that a primitive function of $\sqrt{1+(\beta(x-a))^2}$ with respect to x is given by

$$x \mapsto \frac{1}{2\beta}\left(\beta(x-a)\sqrt{1+(\beta(x-a))^2} + \operatorname{arcsinh}(\beta(x-a))\right), \tag{6.42}$$

in which arcsinh is the inverse hyperbolic sine. It follows that

$$\begin{aligned}\int_{1/3}^1 \sqrt{1+(\beta x)^2}\, dx &= \frac{1}{2\beta}\left(\beta\sqrt{1+\beta^2} + \operatorname{arcsinh}(\beta) - \frac{\beta}{3}\sqrt{1+(\beta/3)^2} - \operatorname{arcsinh}(\beta/3)\right) \\ &= \frac{1}{2}\sqrt{\beta^2+1} - \frac{1}{18}\sqrt{\beta^2+9} + \frac{1}{2\beta}\operatorname{arcsinh}(\beta) - \frac{1}{2\beta}\operatorname{arcsinh}(\beta/3).\end{aligned} \tag{6.43}$$

Similarly, one can derive that

$$\begin{aligned}\int_{1/3}^{1} \sqrt{1+(\beta(x-\tfrac{2}{3}))^2}\,dx &= \int_{-1/3}^{1/3} \sqrt{1+(\beta x)^2}\,dx \\ &= \frac{1}{\beta}\left(\frac{\beta}{3}\sqrt{1+\beta^2/9} + \operatorname{arcsinh}(\beta/3)\right) \\ &= \frac{1}{9}\sqrt{\beta^2+9} + \frac{1}{\beta}\operatorname{arcsinh}(\beta/3), \end{aligned} \tag{6.44}$$

where we use that $\sqrt{1+x^2}$ is an even function. Put together, we obtain

$$\gamma_\beta = \frac{1}{2} + \frac{9}{8}\frac{1}{\beta}\left(\frac{1}{2}\sqrt{\beta^2+1} - \frac{1}{6}\sqrt{\beta^2+9} + \frac{1}{2\beta}\operatorname{arcsinh}(\beta) - \frac{1}{2\beta}\operatorname{arcsinh}(\beta/3)\right) \tag{6.45}$$

Using that $\operatorname{arcsinh}(x) = \log(x+\sqrt{1+x^2})$, we further reduce the above to

$$\gamma_\beta = \frac{1}{2} + \frac{9}{16}\sqrt{1+1/\beta^2} - \frac{3}{16}\sqrt{1+9/\beta^2} + \frac{9}{16\beta^2}\log(3) - \frac{9}{16\beta^2}\log\left(\frac{1+\sqrt{1+9/\beta^2}}{1+\sqrt{1+1/\beta^2}}\right). \tag{6.46}$$

Letting $\beta \longrightarrow \infty$, it follows that $\gamma_\beta \longrightarrow 1/2 + 9/16 - 3/16 = 7/8 = \gamma$, as it should. □

The lemma below provides a crucial bound on the convergence rate of $\gamma - \gamma_\beta$.

Lemma 6.3. *In The Uniform Example it holds that $\gamma - \gamma_\beta < 1/\beta^2$ for all $\beta > 0$.*

Proof. Rewrite the difference $\gamma - \gamma_\beta$ as

$$\gamma - \gamma_\beta = \frac{3}{8} - \frac{3}{16}\left(\sqrt{9+\frac{9}{\beta^2}} - \sqrt{1+\frac{9}{\beta^2}}\right) - \frac{9}{16\beta^2}\log\left(\frac{3+\sqrt{9+9/\beta^2}}{1+\sqrt{1+9/\beta^2}}\right). \tag{6.47}$$

The expression within the logarithm is larger than 1. Hence,

$$\gamma - \gamma_\beta < \frac{3}{8} - \frac{3}{16}\left(\sqrt{9+\frac{9}{\beta^2}} - \sqrt{1+\frac{9}{\beta^2}}\right). \tag{6.48}$$

For $0 < \beta < \sqrt{8/3}$ it holds that $2 - 16/(3\beta^2) < 0$. For $\beta \geq \sqrt{8/3}$, observe that the following expression is strictly positive:

$$9\beta^6 - \frac{1}{4}\beta^6\left(2 - \frac{16}{3\beta^2}\right)^2\left(9 + \frac{9}{\beta^2}\right) = 9\beta^6 - \left(3\beta^2 - 8\right)^2\left(\beta^2 + 1\right) = 39\beta^4 - 16\beta^2 - 64. \tag{6.49}$$

Hence, for all $\beta > 0$, the above shows that

$$\left(2 - \frac{16}{3\beta^2}\right)\sqrt{9 + \frac{9}{\beta^2}} < 6. \tag{6.50}$$

Similarly, it can be shown that

$$\left(2 - \frac{16}{3\beta^2}\right)\sqrt{1 + \frac{9}{\beta^2}} < 2. \tag{6.51}$$

Combining the two bounds results in

$$\left(2 - \tfrac{16}{3\beta^2}\right)\left(\sqrt{9 + \tfrac{9}{\beta^2}} + \sqrt{1 + \tfrac{9}{\beta^2}}\right) < 8 = \left(\sqrt{9 + \tfrac{9}{\beta^2}} + \sqrt{1 + \tfrac{9}{\beta^2}}\right)\left(\sqrt{9 + \tfrac{9}{\beta^2}} - \sqrt{1 + \tfrac{9}{\beta^2}}\right). \tag{6.52}$$

It proves that

$$\sqrt{9 + \frac{9}{\beta^2}} - \sqrt{1 + \frac{9}{\beta^2}} > 2 - \frac{16}{3\beta^2}, \tag{6.53}$$

for all $\beta > 0$, and hence

$$\frac{3}{8} - \frac{3}{16}\left(\sqrt{9 + \frac{9}{\beta^2}} - \sqrt{1 + \frac{9}{\beta^2}}\right) < \frac{1}{\beta^2}. \tag{6.54}$$

This concludes the proof that $\gamma - \gamma_\beta < 1/\beta^2$ for all $\beta > 0$. □

Lemma 6.2 and Lemma 6.3 enable us to prove that $\psi_*^2 - \psi_\beta^2 > -C/\beta^2$ for $C = 15/8$ and all $\beta > 0$.

Theorem 6.3. *In The Uniform Example it holds that*

$$\psi_*^2 - \psi_\beta^2 > -\frac{C}{\beta^2}, \tag{6.55}$$

with $C = 15/8$ *and for all* $\beta > 0$.

Proof. The third variance term ψ_*^2 can be computed directly, yielding

$$\begin{aligned}
\psi_*^2 + \gamma^2 &= \int_0^1 \left(\int_0^1 w_*(x-y)\, dF_Y(y) \right)^2 \, dF_X(x) \\
&= \left(\frac{3}{2}\right)^3 \cdot \int_{1/3}^{2/3} \left(\int_0^x 1\, dy \right)^2 \, dx + \left(\frac{3}{2}\right)^3 \cdot \int_{2/3}^{1} \left(\int_0^{2/3} 1\, dy \right)^2 \, dx \\
&= \left(\frac{3}{2}\right)^3 \cdot \left[\frac{1}{3} x^3 \right]_{1/3}^{2/3} \, dx + \left(\frac{3}{2}\right)^3 \cdot \frac{1}{3} \cdot \left(\frac{2}{3}\right)^2 \\
&= \frac{7}{24} + \frac{1}{2} = \frac{19}{24}.
\end{aligned} \tag{6.56}$$

Hence, using the result from Lemma 6.1, we obtain

$$\psi_*^2 = \frac{19}{24} + \left(\frac{7}{8}\right)^2 = \frac{5}{192}. \tag{6.57}$$

For ψ_β^2, we first derive the lower bound

$$\begin{aligned}
\frac{1}{\beta^2} \int_{1/3}^{1} \sqrt{1 + (\beta x)^2} \sqrt{1 + (\beta(x - \tfrac{2}{3}))^2} \, dx &> \frac{1}{\beta^2} \int_{1/3}^{1} \sqrt{(\beta x)^2} \sqrt{(\beta(x - \tfrac{2}{3}))^2} \, dx \\
&= \int_{1/3}^{2/3} \left(\frac{2}{3} - x\right) x \, dx + \int_{2/3}^{1} \left(x - \frac{2}{3}\right) x \, dx \\
&= \frac{1}{3}\left[x^2 - x^3\right]_{1/3}^{2/3} + \frac{1}{3}\left[x^3 - x^2\right]_{2/3}^{1} \\
&= \frac{2}{27}
\end{aligned} \tag{6.58}$$

Rewrite ψ_β^2 in a convenient way, resulting in

$$\begin{aligned}\psi_\beta^2 &= \int_0^1 \left(\int_0^1 \left(w_\beta(x-y) - \tfrac{1}{2} + \tfrac{1}{2} - \gamma \right)\, dF_Y(y) \right)^2 \, dF_X(x) \\ &= \frac{9}{4} \cdot \int_{1/3}^1 \left(\int_0^{2/3} \left(w_\beta(x-y) - \tfrac{1}{2} \right)\, dy \right)^2 \, dx + 2\left(\frac{1}{2} - \gamma\right)\left(\gamma_\beta - \frac{1}{2}\right) + \left(\frac{1}{2} - \gamma\right)^2 .\end{aligned} \tag{6.59}$$

Then, compute

$$\begin{aligned}&\int_{1/3}^1 \left(\int_0^{2/3} \left(w_\beta(x-y) - \tfrac{1}{2} \right)\, dy \right)^2 \, dx \\ &= \frac{1}{4\beta^2} \int_{1/3}^1 \left(\sqrt{1+(\beta x)^2} - \sqrt{1+(\beta(x-\tfrac{2}{3}))^2} \right)^2 \, dx \\ &= \frac{1}{4\beta^2} \int_{1/3}^1 \left(1 + (\beta x)^2 - 2\sqrt{1+(\beta x)^2}\sqrt{1+(\beta(x-\tfrac{2}{3}))^2} + 1 + (\beta(x-\tfrac{2}{3}))^2 \right) \, dx \\ &= \frac{1}{4\beta^2}\frac{4}{3} + \left[\tfrac{1}{12}x^3\right]_{1/3}^1 + \left[\tfrac{1}{12}\left(x-\tfrac{2}{3}\right)^3\right]_{1/3}^1 - \frac{1}{2\beta^2}\int_{1/3}^1 \sqrt{1+(\beta x)^2}\sqrt{1+(\beta(x-\tfrac{2}{3}))^2}\, dx \\ &= \frac{7}{81} + \frac{1}{3\beta^2} - \frac{1}{2\beta^2}\int_{1/3}^1 \sqrt{1+(\beta x)^2}\sqrt{1+(\beta(x-\tfrac{2}{3}))^2}\, dx \\ &< \frac{1}{6} + \frac{1}{3\beta^2}.\end{aligned} \tag{6.60}$$

The last inequality is where inequality (6.58) is used. Substitute the bound into Equation (6.59) to find

$$\begin{aligned}\psi_\beta^2 &< \frac{9}{4}\left(\frac{1}{6} + \frac{1}{3\beta^2}\right) + 2\left(\frac{1}{2} - \gamma\right)\left(\gamma_\beta - \frac{1}{2}\right) + \left(\frac{1}{2} - \gamma\right)^2 \\ &= \frac{131}{192} + \frac{9}{8\beta^2} - \frac{3}{4}\gamma_\beta \\ &= \frac{5}{192} + \frac{9}{8\beta^2} - \frac{3}{4}(\gamma_\beta - \gamma),\end{aligned} \tag{6.61}$$

which converges to $5/192 = \psi_*^2$ as $\beta \to \infty$, as it should. Finally, Lemma 6.3 implicates that

$$\psi_*^2 - \psi_\beta^2 > \frac{3}{4}(\gamma_\beta - \gamma) - \frac{9}{8\beta^2} > -\frac{15}{8\beta^2}, \tag{6.62}$$

for all $\beta > 0$. This completes the proof. □

The combination of Theorem 6.1, Theorem 6.3 together with Lemma 6.1, and Theorem 6.2 yields the proof of Conjecture 6.1 for The Uniform Example. As clearly stated in Conjecture 6.1, we expect that the same result holds for *more general* distribution functions. Investigating such generalisations is left as a future research direction.

6.5 Implications for model selection

If Conjecture 6.1 is generally true, we believe that it implies that softAUC estimators, for β sufficiently large, are also better model selectors than the standard AUC estimator. As a first step, we prove the following corollary of Conjecture 6.1. The implications for model selection are discussed after the proof.

Corollary 6.1. *Assume that two scoring functions have been trained on the same data set, resulting in true AUCs γ^A and γ^B, and scores S^A and S^B. If Conjecture 6.1 is correct, and S^A and S^B are uncorrelated, then the estimator of the difference $\gamma_A - \gamma_B$ based on η_β has a smaller MSE than that based on η, for β sufficiently large.*

Proof. Let $S^A = \{X_1^A, \ldots, X_n^A, Y_1^A, \ldots, Y_n^A\}$ be the $(n+m)$-vector of the scores of the n positive and m negative labels, corresponding to model A. Define S^B corresponding to model B similarly. The MSE of $\widehat{\eta}(S^A) - \widehat{\eta}(S^B)$ as estimator of $\gamma^A - \gamma^B$ can be written as

$$\begin{aligned}
&\mathbb{E}[\{(\widehat{\eta}(S^A) - \widehat{\eta}(S^B)) - (\gamma^A - \gamma^B)\}^2] \\
&\quad = \mathbb{E}[\{(\widehat{\eta}(S^A) - \gamma^A) - (\widehat{\eta}(S^B) - \gamma^B)\}^2] \\
&\quad = \mathbb{E}[(\widehat{\eta}(S^A) - \gamma^A)^2] + \mathbb{E}[(\widehat{\eta}(S^B) - \gamma^B)^2] - 2\,\mathbb{E}[(\widehat{\eta}(S^A) - \gamma^A)(\widehat{\eta}(S^B) - \gamma^B)] \\
&\quad = M_*^A(n,m) + M_*^B(n,m) - 2\,\mathbb{E}[\widehat{\eta}(S^A) - \gamma^A]\,\mathbb{E}[\widehat{\eta}(S^B) - \gamma^B]. \\
&\quad = M_*^A(n,m) + M_*^B(n,m)
\end{aligned} \tag{6.63}$$

In the third equality, we use that S^A and S^B are uncorrelated. In the fourth equation, we use that $\widehat{\eta}$ is an unbiased estimator of the true AUC. Similarly, we find

$$\begin{aligned}&\mathbb{E}[\{(\widehat{\eta}_\beta(S^A) - \widehat{\eta}_\beta(S^B)) - (\gamma^A - \gamma^B)\}^2] \\ &\qquad = M^A_\beta(n,m) + M^B_\beta(n,m) - 2\,\mathbb{E}[\widehat{\eta}_\beta(S^A) - \gamma^A]\,\mathbb{E}[\widehat{\eta}_\beta(S^B) - \gamma^B)]. \qquad (6.64)\end{aligned}$$

It follows that the difference equals

$$\begin{aligned}&\mathbb{E}[\{(\widehat{\eta}(S^A) - \widehat{\eta}(S^B)) - (\gamma^A - \gamma^B)\}^2] - \mathbb{E}[\{(\widehat{\eta}_\beta(S^A) - \widehat{\eta}_\beta(S^B)) - (\gamma^A - \gamma^B)\}^2] \\ &\qquad = M^A_\beta(n,m) - M^A_* + M^B_\beta(n,m) - M^B_*(n,m) \\ &\qquad + 2\,\mathbb{E}[\widehat{\eta}_\beta(S^A) - \gamma^A]\,\mathbb{E}[\widehat{\eta}_\beta(S^B) - \gamma^B]. \qquad (6.65)\end{aligned}$$

The identity $x^2 + 2xy + y^2 = (x+y)^2$ proves the inequality $2xy \geq -x^2 - y^2$, as $(x+y)^2 \geq 0$. Hence, it holds that

$$\begin{aligned}2\,\mathbb{E}[\widehat{\eta}_\beta(S^A) - \gamma^A]\,\mathbb{E}[\widehat{\eta}_\beta(S^B) - \gamma^B] &\geq -(\mathbb{E}[\widehat{\eta}_\beta(S^A) - \gamma^A])^2 - (\mathbb{E}[\widehat{\eta}_\beta(S^B) - \gamma^B])^2 \\ &\geq -\frac{D}{\beta^2}. \qquad (6.66)\end{aligned}$$

The second inequality follows from applying Theorem 6.2 twice and choosing $D = D^A + D^B$. Finally, if Conjecture 6.1 is correct, it can be applied to both $M^A_* - M^A_\beta$ and $M^B_* - M^B_\beta$ and the result follows. □

In reality, S^A and S^B will most likely have a nonzero (positive) correlation. So, we have to compare the covariance of $\widehat{\eta}$ with that of $\widehat{\eta}_\beta$. We remark (warning) that Theorem 6.1 does not state the same as Conjecture 6.2, but there are some connections. To make the connections clear, define (1) $\Delta_* := \eta(S^A) - \eta(S^B)$ and (2) $\Delta_\beta := \eta(S^A_\beta) - \eta(S^B_\beta)$. If $\gamma^A_\beta - \gamma^B_\beta > \gamma^A - \gamma^B > 0$ (for β sufficiently large), then it is intuitively clear that (a) Δ_β is larger than Δ_* and (b) has a smaller variance as well according to Theorem 6.1. Conversely, if $0 < \gamma^A_\beta - \gamma^B_\beta < \gamma^A - \gamma^B$ (for β sufficiently large), then perhaps the smaller variance of Δ_β can be used to prove Conjecture 6.2. However, this suggestion is left as future work.

6.6 Chapter conclusions

We have provided a thorough theoretical re-evaluation of softAUC estimators, complementing (and, at least partly, contradicting) the empirical evidence in the literature. From the results that we have been able to prove so far, one may already conclude that, in small data sets, softAUC estimators might provide better estimates of the true AUC than the standard AUC estimator. However, tuning the parameter β might be challenging in small data sets as well. So, our main recommendation is to further investigate to what extent our theoretical results generalise to larger data sets.

Supported by the computations in Section 6.4, we believe that the conditions in Conjecture 6.1 are sufficient. Moreover, reflecting on the proofs derived in Section 6.3, we believe that most of the conditions are also necessary. The absolute continuity of F_X and F_Y with respect to the Lebesgue measure on $[0,1]$ might be relaxed, as long as it holds for the measure induced by the CDF G of $X-Y$ around $t=0$, i.e., as long as g (1) exists around $t=0$, (2) is continuous there, and (3) is strictly positive there. One might attempt to construct counterexamples, but we believe that the conditions on g are rather minimal as they are. The condition that f is square integrable on $(-\infty,0]$ is necessary to prove Theorem 6.2; a counterexample was discussed there. The other conditions on f are included to ensure that $w_\beta(t)=f(\beta t)$ converges to $w_*(t)$ as $\beta\to\infty$, although the monotonicity and point-symmetry are not strictly necessary to that end. However, these conditions are rather convenient in deriving our proofs and still allow for sufficient flexibility in the choice of f.

We anticipate that the remaining open problem mentioned in this chapter might be handled by one of the following two attempts. At first, we might extend the results of Section 6.4 to other indicator functions and then to (suitable) step functions. Consequently, we may approximate the distribution functions F_X and F_Y by a sequence of step functions and use either the monotone convergence theorem or the dominated convergence theorem to swap the limits and integrals. Secondly, we could rewrite the difference $\psi_*^2-\psi_\beta^2$ as a single integral with an integrand of the form a^2-b^2 and write that as the product $(a-b)(a+b)$. Then, we might split the domain of integration into points where $x-y$ is close to 0 and points where $|x-y|>c$, where c is some fixed, small constant. Within the first part of the domain, we expect the integral to be positive if we assume that the probability density function g is differentiable and increasing at $t=0$. Within the second part of the domain, we believe that the absolute value of the integral can be bounded from above by a constant multiple of $1/\beta^2$. The desired result would then be produced.

A final remark is that the choice of the algebraic sigmoid function allows for a more thorough empirical study as well. The key is that when using the logistic function, the researcher quickly encounters numeric overflow in the exponential function, whereas the algebraic sigmoid function as used by us can be evaluated accurately and computationally efficiently for much larger values of β. Therefore, the final direction for future research that we recommend is to evaluate the difference $\psi_*^2 - \psi_\beta^2$ for a wide variety of distribution functions (if necessary, at large values of β) employing the algebraic sigmoid function. This recommendation could prove useful in reducing or extending the assumptions that we propose in our conjectures.

APPENDIX

6.A Theoretical derivations for The LDA Example

This appendix contains the proofs of the propositions stated in The LDA Example (see Section 6.1).

Proof of Proposition 6.1. The solutions to the equation $\alpha_1 f_1(x) = \alpha_2 f_2(x)$ are precisely the (affine) hyperplane (see also Hastie et al., 2009) given by

$$x^T \Sigma^{-1}(\mu_2 - \mu_1) - \frac{1}{2}(\mu_2 + \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1) + \log \frac{\alpha_2}{\alpha_1} = 0. \tag{6.67}$$

The region where the left-hand side of the above equation is smaller than 0, i.e., $\alpha_2 f_2(x) < \alpha_1 f_1(x)$, is referred to as region R_1. The other side of the hyperplane is referred to as region R_2. The expectation of the LDA-estimator $\widehat{\alpha}_L$ for $\alpha_1 = \alpha$, using the optimal Bayes classifier, can then be expressed as

$$\mathbb{E}[\widehat{\alpha}_L] = \int_{R_1} \alpha_1 f_1(x) + \alpha_2 f_2(x)\, dx. \tag{6.68}$$

We first consider the situation for $\alpha \neq 0.5$. Assume, without loss of generality, that $\alpha < 0.5$ and hence $\alpha_1 < \alpha_2$. We will show that $\mathbb{E}[\widehat{\alpha}_L] < \alpha$.

The matrix Σ is positive definite, because it is assumed to be a full-rank covariance matrix. We may thus define

$$\sigma := \frac{1}{\sqrt{(\mu_2 - \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1)}}, \tag{6.69}$$

as $(\mu_2 - \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1) > 0$. Consider the affine transformation

$$A : \mathbb{R}^r \to \mathbb{R} : x \mapsto \sigma^2 (x - \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1). \tag{6.70}$$

Observe that A maps μ_1 to 0, μ_2 to 1, and any point on the affine hyperplane given by Equation (6.67) onto the point

$$x_d = \frac{1}{2} - \sigma^2 \log \frac{\alpha_2}{\alpha_1}. \tag{6.71}$$

Next, write $Y_i = AX_i$ where $X_i \sim \mathcal{N}(\mu_i, \Sigma)$, for $i = 1, 2$, and write g_1 and g_2 for the respective densities. It follows that $Y_1 \sim \mathcal{N}(0, \sigma^2)$ and $Y_2 \sim \mathcal{N}(1, \sigma^2)$ and, for $i = 1, 2$, that

$$\int_{R_1} \alpha_i f_i(x)\, dx = \int_{-\infty}^{x_d} \alpha_i g_i(x)\, dx, \tag{6.72}$$

Moreover, from $\alpha_1 < \alpha_2$ it follows that $\alpha_2 g_2(x) > \alpha_1 g_1(x)$ for $x > x_d$ and that $\alpha_2 g_2(x) < \alpha_1 g_1(x)$ for $x < x_d$.

To complete the proof, we distinguish two cases. In the first case, assume that $x_d \leq 0$. It follows that

$$\begin{aligned} \mathbb{E}[\widehat{\alpha}_L] &= \int_{R_1} \alpha_1 f_1(x) + \alpha_2 f_2(x)\, dx = \int_{-\infty}^{x_d} \alpha_1 g_1(x) + \alpha_2 g_2(x)\, dx \\ &< \int_{-\infty}^{x_d} 2\alpha_1 g_1(x)\, dx < \int_{-\infty}^{0} 2\alpha_1 g_1(x)\, dx = \alpha_1 = \alpha. \end{aligned} \tag{6.73}$$

In the second case, assume that $x_d > 0$, but note that we must have $x_d < \frac{1}{2}$, because $\alpha_1 < \alpha_2$. Define $Y_3 \sim \mathcal{N}(2x_d - 1, \sigma^2)$ having density $g_3(x) = g_2(2x_d - x)$. Because $g_3(x_d) = g_2(x_d)$, it follows that $\alpha_1 g_1(x) = \alpha_2 g_3(x)$ if and only if $x = x_d$. Moreover, as $2x_d - 1 < 0$, we find

$$\alpha_2 g_3(2x_d - 1) = \alpha_2 g_2(1) = \frac{\alpha_2}{\sqrt{2\pi\sigma^2}} > \frac{\alpha_1}{\sqrt{2\pi\sigma^2}} = \alpha_1 g_1(0) > \alpha_1 g_1(2x_d - 1). \tag{6.74}$$

Hence, $\alpha_2 g_3(x) > \alpha_1 g_1(x)$ for all $x < x_d$ and $\alpha_2 g_3(x) < \alpha_1 g_1(x)$ for all $x > x_d$. It follows that

$$\int_{-\infty}^{x_d} \alpha_2 g_2(x)\, dx = \int_{x_d}^{\infty} \alpha_2 g_3(x)\, dx < \int_{x_d}^{\infty} \alpha_1 g_1(x)\, dx. \tag{6.75}$$

We conclude that

$$\mathbb{E}[\widehat{\alpha}_L] = \int_{-\infty}^{x_d} \alpha_1 g_1(x) + \alpha_2 g_2(x) < \int_{-\infty}^{\infty} \alpha_1 g_1(x)\, dx = \alpha_1 = \alpha. \tag{6.76}$$

Finally, if $\alpha = 0.5$ and hence $\alpha_1 = \alpha_2$, then in the above $x_d = \frac{1}{2}$ and $\alpha_2 g_3(x) = \alpha_1 g_1(x)$ for any $x \in \mathbb{R}$. It then follows that

$$\int_{-\infty}^{x_d} \alpha_1 g_1(x) + \alpha_2 g_2(x)\, dx = \alpha_1 = \alpha. \tag{6.77}$$

Thus, $\mathbb{E}[\widehat{\alpha}_L] = \alpha$. This concludes the proof of Proposition 6.1. □

The following (rather trivial) result shows that aggregating scores prevents misclassification bias if all modelling assumptions are satisfied and $\alpha^M = \alpha$.

Lemma 6.4. *Assume that $\alpha^M = \alpha$. Let X be a random vector in the feature space $\mathcal{X}$ that is drawn from f_1 with probability $\alpha_1 = \alpha$ and from f_2 with probability $\alpha_2 = 1 - \alpha$. Denote the probability that a realisation x of X is classified into class $k \in \{1, 2\}$ by $p_k(x)$. The expectation $\mathbb{E}_X[p_1]$ then equals the base rate α.*

Proof. The probability density function f_X of X is given by

$$f_X(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x). \tag{6.78}$$

The score $p_1(x)$ estimated by LDA is given by

$$p_1(x) = \frac{\alpha_1^M f_1(x)}{\alpha_1^M f_1(x) + \alpha_2^M f_2(x)} = \frac{\alpha_1^M f_1(x)}{f_X(x)}, \tag{6.79}$$

using that $\boldsymbol{\alpha}^M = \boldsymbol{\alpha}$ in the last equality. It directly follows that

$$\mathbb{E}_X[p_1] = \int_{x \in \mathcal{X}} \frac{\alpha_1^M f_1(x)}{f_X(x)}\, dF_X(x) = \int_{x \in \mathcal{X}} \alpha_1^M f_1(x)\, dx = \alpha_1^M = \alpha_1 = \alpha. \tag{6.80}$$

This concludes the proof. □

Finally, we provide the proof of Proposition 6.2.

Proof of Proposition 6.2. As in the proof of Lemma 6.4, the probability density function f_X of X is given by

$$f_X(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x). \tag{6.81}$$

The score $p_1(x)$ estimated by LDA is (again) given by

$$p_1(x) = \frac{\alpha_1^M f_1(x)}{\alpha_1^M f_1(x) + \alpha_2^M f_2(x)}. \tag{6.82}$$

Define the unbiased probability $p_1^U(x)$ as

$$p_1^U(x) = \frac{\alpha_1 f_1(x)}{\alpha_1 f_1(x) + \alpha_2 f_2(x)}. \tag{6.83}$$

First, assume that $\alpha_1^M > \alpha_1$. Compare $p_1(x)$ and $p_1^U(x)$ by viewing

$$\begin{aligned}(\alpha_1 f_1 + \alpha_2 f_2)(\alpha_1^M f_1 + \alpha_2^M f_2)(p_1 - p_1^U) &= \alpha_1^M f_1(\alpha_1 f_1 + \alpha_2 f_2) - \alpha_1 f_1(\alpha_1^M f_1 + \alpha_2^M f_2)\\ &= \alpha_1^M \alpha_2 f_1 f_2 - \alpha_1 \alpha_2^M f_1 f_2\\ &= (\alpha_1^M \alpha_2 - \alpha_1 \alpha_2^M) f_1 f_2\\ &= (\alpha_1^M - \alpha_1) f_1 f_2 > 0. \end{aligned} \tag{6.84}$$

The derivation shows that $p_1(x) > p_1^U(x)$ for any $x \in \mathcal{X}$. It implies that aggregating probability vectors still results in a biased estimate for α:

$$\begin{aligned}\mathbb{E}_{\mathcal{X}}[p_1] = \int_{\mathcal{X}} p_1(x)\, dF_X(x) &> \int_{\mathcal{X}} p_1^U(x)\, dF_X(x)\\ &= \int_{\mathcal{X}} \frac{\alpha_1 f_1(x)}{f_X(x)} f_X(x)\, dx = \alpha_1. \end{aligned} \tag{6.85}$$

Similarly, if $\alpha_1^M < \alpha_1$, we will find $\mathbb{E}_{\mathcal{X}}[p_1] < \alpha_1$. This concludes the proof of Proposition 6.2. □

The proof of Proposition 6.2 highlights a fundamental issue: if the base rate differs between the training data and the unlabelled data, then the score estimated by LDA is biased *for every single data point*. Moreover, the bias is *in the same direction for all data points*. The implication is that the bias will never cancel out, neither when aggregating scores over the entire population, nor when aggregating them over any subpopulation.

CHAPTER 7

CONCLUSIONS

In this chapter we first answer the four research questions in Section 7.1. We then answer the problem statement and arrive at the main conclusion of this thesis in Section 7.2. Finally, we recommend five directions for future research in Section 7.3.

7.1 Answers to the research questions

Below, we reiterate the research questions as introduced in Chapter 1 and provide an answer to each of them separately.

> **RQ1**: *Which estimator of the base rate, in particular when dealing with concept drift, has the smallest MSE in finite populations?*

As announced in Section 1.6, we answered RQ1 in two steps, namely under two different assumptions (A1 and A2). We recall that assumption A1 corresponds to the double sampling scheme and that assumption A2 corresponds to a specific type of concept drift called prior probability shift (see Section 1.6).

In Chapter 2, we answered RQ1 under assumption A1. We derived analytic expressions for the MSE of both the misclassification estimator $\hat{\alpha}_p$ and the calibration estimator $\hat{\alpha}_c$ up to and including terms of order $1/n$, where n equals the sample size of the test set. We then used these expressions to prove theoretically that the MSE of the calibration estimator is smaller than that of the misclassification estimator of any base rate and any classification algorithm.

In Chapter 3, we derived similar analytic expressions, but now under assumption A2. Based on a numerical comparison of the resulting expressions, we were able to show how (the sign of) the difference $D(\hat{\alpha}_p, \hat{\alpha}_c)$ of the MSEs of the two estimators depends on (a) the level of drift δ, (b) the initial base rate α, (c) the sample size n of the test set, and (d) the performance of the classifier in terms of p_{00} and p_{11}. The conclusion of Chapter 3 is that the MSE of the calibration estimator is smaller than that of the misclassification estimator only when the performance of the classifier is low or when the drift is close to 0. Therefore, our recommendation is that the calibration estimator should *not* be applied to data streams or time series data, unless training and test data in each time period are available to (i) retrain the classifier and hence (ii) adapt to concept drift.

RQ2: *How can we leverage identification regions of misclassification probabilities in order to reduce the MSE of classifier-based statistics even further?*

In Chapter 4, we derived the posterior distribution (for conjugate priors) of the model parameters used when employing the misclassification estimator. We then proposed a new Bayesian method to correct for misclassification bias. The method that we proposed is to use the misclassification estimator and impose constraints on the prior distribution of the model parameters, leveraging their identification regions. We argued that our method is successful when the sample size n of the test set is much smaller than the population size N of the unlabelled data set. By means of a simulation study, we showed that our method reduces the MSE of the standard misclassification estimator, indeed in particular when dealing with small test sets. Hence, the method that we developed in Chapter 4 provides an answer to RQ2.

RQ3: *To what extent can statistical learning be used to improve the accuracy of estimates of cross-border Internet purchases within the EU?*

In Chapter 5, we first discovered that existing consumer-survey approaches led to a serious underestimation of cross-border Internet purchases within the EU. We argued that language plays a pivotal role. We therefore identified three *supply-side* data sources that contain information on cross-border Internet purchases within the EU. We then developed a general-purpose approach for firm-level record linkage to combine the data sources. The approach is based on approximate string matching, locality-sensitive hashing, and statistical learning (i.e., support vector classification with radial basis function kernel). Furthermore, we combined that approach with web scraping techniques and statistical learning (i.e., random forest) to develop a new data-driven approach to estimate cross-border Internet purchases within the EU, which is internationally consistent and comparable. Finally, we applied our approach to data from the Netherlands for the year 2016. It resulted in an estimate of cross-border Internet purchases that is six times as high as existing estimates, having a standard error of only 8%. Thus, the answer to RQ3 is that statistical learning can improve the accuracy of estimates of cross-border Internet purchases within the EU significantly.

7.2 Answer to the problem statement

We are now able to provide an answer to the problem statement based on the answers to the research questions.

> **Problem statement (PS)**: *In what way can we reduce misclassification bias in statistical learning so that we obtain more accurate classifier-based statistics?*

Our answer to the problem statement contains a theoretical component (based on the answers to RQ1 and RQ2) and an empirical component (based on the answer to RQ3).

Theoretical component. Reducing misclassification bias might simultaneously increase the variance of classifier-based statistics (see Fig. 1.1). Therefore, we investigated the *MSE* of two popular bias correction methods: (1) the misclassification estimator and (2) the calibration estimator. Among these two estimators, the calibration estimator has the smallest MSE *if training and test data are available in each time period.* However, if training and test data are scarce, the misclassification estimator often has a smaller MSE. The MSE can be reduced even further by imposing parameter constraints.

Empirical component. Our results from Chapter 5 show that statistical learning has the potential to improve official statistics significantly. Moreover, even imperfect classification algorithms can be used to obtain accurate classifier-based statistics, as long as we correct for misclassification bias.

Thus, we may conclude that misclassification bias in statistical learning can be handled adequately. Therefore, statistical learning can indeed further the field of official statistics. Our recommendation is to apply statistical learning either to develop new official statistics or to improve existing official statistics. Based on the results from Chapter 3, we stress that statistical learning cannot replace domain experts. Hence, prudence is recommended when evaluating the cost efficiency of implementing statistical learning methods in official statistics.

7.3 Future work

Below, we recommend five directions for future research to further the methodological understanding of statistical learning. We believe that devoting attention to these five research directions will improve applications of statistical learning methods in official statistics.

First, we support the recommendation by González et al. (2017) that the theoretical properties of (the more advanced) methods from quantification learning should be investigated. More specifically, inspired by González et al. (2017), we distinguish three categories of methods to examine. The first category of methods that we distinguish is *model averaging* of classifier-based quantifiers. It is well known that model averaging reduces the MSE, and we believe that a theoretical examination is still tractable when averaging the methods that are presented in this thesis. The second category of methods to investigate is the group of methods based on *rankers instead of classifiers*. In fact, many classifiers are a composition of a ranker with a threshold function. The threshold function discards information that is valuable in estimating the base rate. This is shown by the empirical evidence provided by Forman (2006) for his method called median sweep. We believe that the theoretical results presented in Chapter 6 might support the theoretical analysis of quantifiers based on rankers, which we might call ranker-based statistics. The third category of methods is *distribution matching*. We believe that a promising starting point would be to embed the literature on quantification learning methods that are based on distribution matching within the statistical literature on kernel density estimation. There is a rich statistical literature on kernel density estimation (see, e.g., Gramacki, 2018) that might prove to be very useful in the context of quantification learning.

Second, investigating more realistic measurement error models is an important direction for future research. Throughout this thesis, we have assumed the strongest type of nondifferential misclassification, namely independent and identically distributed misclassifications (conditional on the true class). Initially, our work can be extended to misclassifications that are nondifferential within strata of the data only, as considered by Van Delden et al. (2016). We believe that our theoretical results will generalise to that setting rather easily. Thereafter, more complex measurement error models, e.g., those outlined by Schennach (2016), should be investigated. To deal with these more complex measurement error models adequately, we anticipate that more advanced statistical theory needs to be developed.

Third, future work could also focus on other types of aggregation. In this thesis we have studied proportions of a random variable only. A natural extension would be to consider *ratios of random variables*, including growth rates of a random variable over time. A second possible extension is *deaggregation*, either by aggregating over subpopulations instead of the entire target population or by further specifying a dichotomous classification into further subclasses. How can we train a binary quantifier for the entire target population, such that it is also an accurate quantifier for subpopulations or for more detailed classifications? Early results by Scholtus and Van Delden (2020) provide empirical evidence that this is a nontrivial task.

Fourth, future research could address other types of concept drift, extending our results on prior probability shift. We believe that the following three steps should be taken in this direction. A first step should be to generalise the results in Chapter 2 to the broader definition of prior probability shift (called *class drift*, see Webb et al., 2016). A second step should be to investigate *covariate drift*, which occurs whenever $P(X)$ changes over time. Again, we recommend to investigate the restrictive definition of (pure) covariate shift (see Moreno-Torres et al., 2012) before looking into the broader definition of covariate drift by Webb et al. (2016). A third step should be to look into more specific properties of the two types of concept drift just mentioned, for example the *drift frequency, duration and magnitude* (Webb et al., 2016). These three steps will require new theoretical findings and derivations. They will be more intricate than our findings and derivations in

Chapters 2 and 3. However, our results provide a starting point.

A fifth possible direction for future research is to examine the biases that occur in big data sources (Baeza-Yates, 2018; Mehrabi et al., 2019), with selectivity being the most relevant one within the context of official statistics (De Broe et al., 2020). The setting of prior probability shift partially resembles selectivity, but more general settings should be investigated. We believe that a deeper understanding of selectivity in big data sources, complemented with the other four research directions outlined above, will support national statistical institutes to keep providing the detailed and highly accurate statistical information that society demands.

REFERENCES

Agresti, A. & Hitchcock, D. B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, *14*(3), 297–330. doi: 10.1007/s10260-005-0121-y. Cited on pages 11, 86, and 90.

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B. & Salakoski, T. (2009). A comparison of AUC estimators in small-sample studies. In S. Džeroski, P. Guerts & J. Rousu (Eds.), *Proceedings of the 3rd International Workshop on Machine Learning in Systems Biology (MLSB)* (pp. 3–13). Ljubljana. Cited on page 152.

Autor, D., Dorn, D., Hanson, G. H., Pisano, G. & Shu, P. (2017). *Foreign competition and domestic innovation: Evidence from U.S. patents* (Working Paper No. 22879). Cambridge, MA: National Bureau of Economic Research. doi: 10.3386/w22879. Cited on pages 107 and 108.

Baeza-Yates, R. A. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. doi: 10.1145/3209581. Cited on pages 4 and 186.

Bailey, J. P., Gao, G., Jank, W., Lin, M., Lucas, H. C. & Viswanathan, S. (2008). *The long tail is longer than you think: The surprisingly large extent of online sales by small volume sellers.* (Available at SSRN.) doi: 10.2139/ssrn.1132723. Cited on page 111.

Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., ... Fleming, L. (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics & Management Strategy*, *27*(3), 535–553. doi: 10.1111/jems.12259. Cited on page 107.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*(4), 387–415. doi: 10.1016/0022-2496(75)90001-2. Cited on pages 151 and 154.

Bawa, M., Condie, T. & Ganesan, P. (2005). LSH Forest: Self-tuning indexes for similarity search. In A. Ellis & T. Hagino (Eds.), *Proceedings of the 14th International Conference on World Wide Web (WWW)* (pp. 651–660). Chiba. doi: 10.1145/1060745.1060840. Cited on pages 12, 118, and 140.

Beck, M., Dumpert, F. & Feuerhake, J. (2018). Machine learning in official statistics. *arXiv/1812.10422*. Cited on pages 4 and 57.

Bena, J., Ferreira, M. A., Matos, P. & Pires, P. (2017). Are foreign investors locusts? The long-term effects of foreign institutional ownership. *Journal of Financial Economics*, *126*(1), 122–146. doi: 10.1016/j.jfineco.2017.07.005. Cited on page 107.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, *45*(250), 164–180. doi: 10.1080/01621459.1950.10483349. Cited on page 8.

Birnbaum, Z. W. & Klose, O. M. (1957). Bounds on the variance of the Mann-Whitney statistic. *The Annals of Mathematical Statistics*, *28*(4), 933–945. Cited on page 152.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. Cited on page 89.

Blazquez, D., Domenech, J., Gil, J. A. & Pont, A. (2018). Monitoring e-commerce adoption from online data. *Knowledge and Information Systems*, *60*. doi: 10.1007/s10115-018-1233-7. Cited on page 108.

Braaksma, B. & Zeelenberg, C. (2015). “Re-make/Re-model”: Should big data change the modelling paradigm in official statistics? *Statistical Journal of the IAOS*, *31*(2), 193–202. doi: 10.3233/sji-150892. Cited on pages 1 and 3.

Breiman, L. & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, *60*(3), 291–319. doi: 10.2307/1403680. Cited on page 120.

Broder, A. Z. (1997). On the resemblance and containment of documents. In B. Carpentieri, A. De Santis, U. Vaccaro & J. A. Storer (Eds.), *Proceedings of Compression and Complexity of Sequences (SEQUENCES)* (pp. 21–29). Salerno. doi: 10.1109/SEQUEN.1997.666900. Cited on pages 12, 118, and 139.

Bross, I. D. J. (1954). Misclassification in 2 × 2 tables. *Biometrics*, *10*(4), 478–486. doi: 10.2307/3001619. Cited on pages 7, 57, 60, and 85.

Buelens, B., Boonstra, H. J., van den Brakel, J. A. & Daas, P. J. H. (2012). *Shifting paradigms in official statistics: From design-based to model-based to algorithmic inference* (Discussion Paper No. 201218). The Hague/Heerlen: Statistics Netherlands. Cited on page 3.

Buelens, B., de Wolf, P.-P. & Zeelenberg, C. (2016). Model based estimation at Statistics Netherlands. In *Proceedings of the 2016 European Conference on Quality in Official Statistics (Q2016).* Madrid. Retrieved from `https://www.ine.es/q2016/docs/q2016Final00196.pdf.` (Accessed January 2021.) Cited on pages 3, 4, 9, 57, and 59.

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications.* Chapman & Hall/CRC. Cited on pages 7, 8, 9, 19, 25, 26, 57, 58, and 61.

Burger, J., Van Delden, A. & Scholtus, S. (2015). Sensitivity of mixed-source statistics to classification errors. *Journal of Official Statistics*, *31*(3), 489–506. doi: 10.1515/jos-2015-0029. Cited on page 23.

Cardona, M. & Duch-Brown, N. (2016). *Delivery costs and cross-border e-commerce in the EU digital single market* (JRC Working Papers on Digital Economy No. 2016-03). Seville: Joint Research Centre. Cited on page 103.

Cawley, G. C. & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107. Cited on page 86.

Chessa, A. G. (2016). *Processing scanner data in the Dutch CPI: A new methodology and first experiences. Presented at the 2016 UNECE Meeting of the Group of Experts on Consumer Price Indices. Geneva.* `https://pdfs.semanticscholar.org/a9d6/5dde47400c041022d5558da855589ef3a23a.pdf`. (Accessed January 2021.) Cited on page 3.

Cohen, W., Ravikumar, P. & Fienberg, S. (2003). A comparison of string metrics for matching names and records. In L. A. Getoor & T. E. Senator (Eds.), *Proceedings of the Workshop on Data Cleaning and Object Consolidation at the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* Washington, D.C.. Cited on pages 12, 117, and 141.

Costa, H., Almeida, D., Vala, F., Marcelino, F. & Caetano, M. (2018). Land cover mapping from remotely sensed and auxiliary data for harmonized official statistics. *ISPRS International Journal of Geo-Information*, *7*(4), 157. doi: 10.3390/ijgi7040157. Cited on pages 81, 84, and 86.

Curier, R. L., De Jong, T. J. A., Strauch, K., Cramer, K., Rosenski, N., Schartner, C., . . . Bromuri, S. (2018). Monitoring spatial sustainable development: Semi-automated analysis of satellite and aerial images for energy transition and sustainability indicators. *arXiv/1810.04881*. Cited on pages 19, 20, and 60.

Czaplewski, R. L. (1992). Misclassification bias in areal estimates. *Photogrammetric Engineering and Remote Sensing*, *58*(2), 189–192. Cited on pages 5, 8, 19, and 21.

Daas, P. J. H., Puts, M. J., Buelens, B. & Van den Hurk, P. A. M. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, *31*(2), 249–262. doi: 10.1515/JOS-2015-0016. Cited on pages 3 and 81.

Davis, J. & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In W. W. Cohen & A. Moore (Eds.), *Proceedings of the 23rd international conference on Machine learning (ICML)* (pp. 233–240). New York, NY. doi: 10.1145/1143844.1143874. Cited on pages 121 and 151.

De Broe, S. M. M. G., Struijs, P., Daas, P. J. H., Van Delden, A., Burger, J., Van den Brakel, J. A., . . . Ypma, W. F. H. (2020). *Updating the paradigm of official statistics* (CBDS Working Paper No. 02-20). The Hague/Heerlen: Statistics Netherlands. Cited on pages 1, 3, 4, 57, and 186.

European Commission. (2009). Regulation of European Statistics. *Official Journal of the European Union*, *L87*, 164–173. `http://data.europa.eu/eli/reg/2009/223/oj`. (Accessed January 2021.) Cited on pages 3 and 57.

European Commission. (2015). *Monitoring the digital economy & society 2016-2021.* `http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=13706`. (Accessed January 2021.) Cited on page 103.

Eurostat. (2017). *European Statistics Code of Practice.* `https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142`. Luxembourg: Publications Office of the European Union. (Accessed January 2021.) doi: 10.2785/914491. Cited on pages 3 and 57.

Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. doi: 10.2307/2286061. Cited on pages 12 and 107.

Forman, G. (2005). Counting positives accurately despite inaccurate classification. In J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge & L. Torgo (Eds.), *Proceedings of the 16th European Conference on Machine Learning (ECML)* (pp. 564–575). Porto: Springer. doi: 10.1007/11564096_55. Cited on pages 5, 57, 109, and 124.

Forman, G. (2006). Quantifying trends accurately despite classifier error and class imbalance. In L. H. Ungar, M. Craven, D. Gunopulos & R. Eliassi-Rad (Eds.), *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)* (pp. 157–166). Philadelphia, PA. doi: 10.1145/1150402.1150423. Cited on pages 149 and 184.

Forman, G. (2008). Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, *17*(2), 164–206. doi: 10.1007/s10618-008-0097-y. Cited on page 5.

Gaba, A. & Winkler, R. L. (1992). Implications of errors in survey data: A Bayesian model. *Management Science*, *38*(7), 913–925. doi: 10.1287/mnsc.38.7.913. Cited on page 9.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), 1–37. doi: 10.1145/2523813. Cited on pages 4, 58, and 59.

Garcia-Bernardo, J. & Takes, F. W. (2018). The effects of data quality on the analysis of corporate board interlock networks. *Information Systems*, *78*, 164–172. doi: 10.1016/j.is.2017.10.005. Cited on page 113.

Goldenberg, I. & Webb, G. I. (2019). Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowledge and Information Systems*, *60*(2), 591–615. doi: 10.1007/s10115-018-1257-z. Cited on page 59.

Goldstein, N. D., Burstyn, I., Newbern, E. C., Tabb, L. P., Gutowski, J. & Welles, S. L. (2016). Bayesian correction of misclassification of pertussis in vaccine effectiveness studies: How much does underreporting matter? *American Journal of Epidemiology*, *183*(11), 1063–1070. doi: 10.1093/aje/kwv273. Cited on pages 10 and 85.

Gomez-Herrera, E., Martens, B. & Turlea, G. (2014). The drivers and impediments for cross-border e-commerce in the EU. *Information Economics and Policy*, *28*, 83–96. doi: 10.1016/j.infoecopol.2014.05.002. Cited on page 103.

González, P., Castaño, A., Chawla, N. V. & del Coz, J. J. (2017). A review on quantification learning. *ACM Computing Surveys*, *50*(5), 74:1–74:40. doi: 10.1145/3117807. Cited on pages 5, 7, 19, 23, 57, 86, 106, and 184.

Gramacki, A. (2018). *Nonparametric kernel density estimation and its computational aspects*. Springer. doi: 10.1007/978-3-319-71688-6. Cited on page 185.

Grassia, A. & Sundberg, R. (1982). Statistical precision in the calibration and use of sorting machines and other classifiers. *Technometrics*, *24*(2), 117–121. doi: 10.1080/00401706.1982.10487732. Cited on page 25.

Graybill, F. A. (1983). *Matrices with applications in statistics* (2nd ed.). Wadsworth. Cited on page 91.

Greenland, S. (2014). Sensitivity analysis and bias analysis. In W. Ahrens & I. Pigeot (Eds.), *Handbook of epidemiology* (p. 685-706). New York, NY: Springer. doi: 10.1007/978-0-387-09834-0_60. Cited on page 19.

Gustafson, P. (2004). *Measurement error and misclassification in statistics and epidemiology: Impact and Bayesian adjustments*. Chapman & Hall/CRC. doi: 10.1201/9780203502761. Cited on pages 10 and 85.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. doi: 10.1016/j.eswa.2016.12.035. Cited on page 86.

Hall, B. H., Jaffe, A. B. & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools* (Working Paper No. 8498). Cambridge, MA: National Bureau of Economic Research. doi: 10.3386/w8498. Cited on page 107.

Han, J., Kamber, M. & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufman. doi: 10.1016/C2009-0-61819-5. Cited on page 120.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer. doi: 10.1007/978-0-387-84858-7. Cited on pages 2, 3, 120, 150, and 175.

Helmbold, D. P. & Long, P. M. (1994). Tracking drifting concepts by minimizing disagreements. *Machine Learning*, *14*(1), 27–45. doi: 10.1007/BF00993161. Cited on page 58.

Hogg, R. V., McKean, J. W. & Craig, A. T. (2018). *Introduction to mathematical statistics* (8th ed.). Pearson. Cited on page 90.

Hopkins, D. J. & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247. doi: 10.1111/j.1540-5907.2009.00428.x. Cited on page 19.

Jaidka, K., Ahmed, S., Skoric, M. & Hilbert, M. (2018). Predicting elections from social media: A three-country, three-method comparative study. *Asian Journal of Communication*, *29*(3), 252-273. doi: 10.1080/01292986.2018.1453849. Cited on page 6.

Jeni, L. A., Cohn, J. F. & De La Torre, F. (2013). Facing imbalanced data – Recommendations for the use of performance metrics. In A. Nijholt, M. Pantic & S. D'Mello (Eds.), *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 245–251). Geneva. doi: 10.1109/ACII.2013.47. Cited on page 121.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, *31*(3), 471–481. doi: 10.3233/SJI-150906. Cited on page 4.

Kloos, K., Meertens, Q. A., Scholtus, S. & Karch, J. D. (2020). Comparing correction methods to reduce misclassification bias. In L. Cao, W. A. Kosters & J. Lijffijt (Eds.), *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC)* (pp. 103–129). Leiden. Cited on pages 58, 59, 60, 61, 62, 63, 64, 67, and 68.

Knottnerus, P. (2003). *Sample survey theory: Some Pythagorean perspectives*. New York, NY: Springer-Verlag. doi: 10.1007/978-0-387-21764-2. Cited on page 38.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference of Artificial Intelligence (IJCAI)* (pp. 1137–1145). Montreal. Cited on page 120.

Kuha, J. & Skinner, C. J. (1997). Categorical data analysis and misclassification. In L. E. Lyberg et al. (Eds.), *Survey measurement and process quality* (pp. 633–670). New York, NY: Wiley. Cited on pages 8, 12, 19, 21, 26, and 57.

Lash, T. L., Fox, M. P. & Fink, A. K. (2009). *Applying quantitative bias analysis to epidemiologic data*. Springer. doi: 10.1007/978-0-387-87959-8. Cited on pages 85 and 109.

Lenstra, A. (2005). Cramér-Rao revisited. *Bernoulli*, *11*(2), 263–282. doi: 10.3150/bj/1116340294. Cited on page 153.

Leskovec, J., Rajaraman, A. & Ullman, J. D. (2014). *Mining of massive datasets* (2nd ed.). Cambridge University Press. doi: 10.1017/CBO9781139924801. Cited on pages 118, 139, and 141.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, *11*(1), 22–31. Cited on pages 117 and 137.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J. & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346–2363. doi: 10.1109/TKDE.2018.2876857. Cited on page 59.

Löw, F., Knöfel, P. & Conrad, C. (2015). Analysis of uncertainty in multi-temporal object-based classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *105*, 91–106. doi: 10.1016/j.isprsjprs.2015.03.004. Cited on pages 19, 86, and 109.

Ma, L., Li, M., Ma, X., Cheng, L., Du, P. & Liu, Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 277–293. doi: 10.1016/j.isprsjprs.2017.06.001. Cited on page 81.

Ma, S., Chai, Y. & Zhang, H. (2018). Rise of cross-border e-commerce exports in China. *China & World Economy*, *26*(3), 63–87. doi: 10.1111/cwe.12243. Cited on page 113.

MacFeely, S. (2016). The continuing evolution of official statistics: Some challenges and opportunities. *Journal of Official Statistics*, *32*(4), 789–810. doi: 10.1515/jos-2016-0041. Cited on page 4.

Majnik, M. & Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intelligent Data Analysis*, *17*(3), 531–558. doi: 10.3233/IDA-130592. Cited on page 151.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. Cited on pages 151 and 154.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. doi: 10.1017/CBO9780511809071. Cited on pages 117 and 122.

Marcus, J. S. & Petropoulos, G. (2016). *E-commerce in Europe: Parcel delivery prices in a digital single market* (Policy Contribution No. 2016-09). Brussels: Bruegel. Cited on page 103.

Martikainen, E., Schmiedel, H. & Takalo, T. (2015). Convergence of European retail payments. *Journal of Banking & Finance*, *50*, 81–91. doi: 10.1016/j.jbankfin.2014.09.021. Cited on page 103.

Meertens, Q. A., Diks, C. G. H., van den Herik, H. J. & Takes, F. W. (2018). A data-driven supply-side approach for measuring cross-border internet purchases. *arXiv/1805.06930*. Cited on pages 96 and 97.

Meertens, Q. A., Diks, C. G. H., Van den Herik, H. J. & Takes, F. W. (2019). A Bayesian approach for accurate classification-based aggregates. In T. Y. Berger-Wolf & N. V. Chawla (Eds.), *Proceedings of the 19th SIAM International Conference on Data Mining (SDM)* (pp. 306–314). Calgary. doi: 10.1137/1.9781611975673.35. Cited on page 33.

Meertens, Q. A., Diks, C. G. H., Van den Herik, H. J. & Takes, F. W. (2020). 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. doi: 10.1111/rssa.12487. Cited on pages 19 and 96.

Meertens, Q. A., Van Delden, A., Scholtus, S. & Takes, F. W. (2019). Bias correction for predicting election outcomes with social media data. In *Extended Abstracts of the 5th International Conference on Computational Social Science (IC2S2)*. Amsterdam. Retrieved from https://www.researchgate.net/publication/333661444_Bias_Correction_for_Predicting_Election_Outcomes_with_Social_Media_Data. (Accessed January 2021.) Cited on page 6.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv/1908.09635*. Cited on pages 4 and 186.

Minges, M. (2016). *In search of cross-border e-commerce trade data* (No. 6). http://unctad.org/en/PublicationsLibrary/tn_unctad_ict4d06_en.pdf. Geneva. (Accessed January 2021.) Cited on page 104.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, *144*(1), 81–117. doi: 10.1016/j.jeconom.2007.12.003. Cited on pages 10, 12, and 93.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530. doi: 10.1016/j.patcog.2011.06.019. Cited on pages 10, 34, 58, 62, and 185.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In M. A. Hearst (Ed.), *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)* (pp. 122–129). Washington, D.C.. Cited on pages 5, 6, and 23.

OECD. (2011). *Quality Framework for OECD Statistical Activities.* https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm. (Accessed January 2021.) Cited on pages 3 and 57.

Oestreicher-Singer, G. & Sundararajan, A. (2012). Recommendation networks and the long tail of electronic commerce. *MIS Quarterly*, *36*(1), 65–83. doi: 10.2307/41410406. Cited on page 111.

Olofsson, P., Foody, G. M., Stehman, S. V. & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, *129*, 122–131. doi: 10.1016/j.rse.2012.10.031. Cited on page 86.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Cited on pages 3 and 120.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. Cited on pages 117 and 137.

Ravi, K. & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. doi: 10.1016/j.knosys.2015.06.015. Cited on page 81.

Ribeiro, S. P., Menghinello, S. & De Backer, K. (2010). *The OECD ORBIS database: Responding to the need for firm-level micro-data in the OECD* (OECD Statistics Working Papers No. 2010-1). Paris: OECD. Cited on page 113.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, *8*(1), 341–377. doi: 10.1146/annurev-economics-080315-015058. Cited on page 185.

Schlimmer, J. C. & Granger, R. H. (1986). Incremental learning from noisy data. *Machine Learning*, *1*(3), 317–354. doi: 10.1007/BF00116895. Cited on page 58.

Scholtus, S. & Van Delden, A. (2020). *The accuracy of estimators based on a binary classifier* (Discussion Paper No. 202007). The Hague: Statistics Netherlands. Cited on pages 6, 7, 19, 23, 24, and 185.

Schu, M. & Morschett, D. (2017). Foreign market selection of online retailers – A path-dependent perspective on influence factors. *International Business Review*, *26*(4), 710-723. doi: 10.1016/j.ibusrev.2017.01.001. Cited on page 103.

Schwartz, J. E. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods & Research*, *13*(4), 435–466. doi: 10.1177/0049124185013004001. Cited on pages 5, 19, and 57.

Sebastiani, F. (2020). Evaluation measures for quantification: An axiomatic approach. *Information Retrieval Journal*, *23*(3), 255–288. doi: 10.1007/s10791-019-09363-y. Cited on page 9.

Sodomka, E., Lahaie, S. & Hillard, D. (2013). A predictive model for advertiser value-per-click in sponsored search. In R. A. Baeza-Yates & S. Moon (Eds.), *Proceedings of the 22nd International Conference on World Wide Web (WWW)* (pp. 1179–1190). Rio de Janeiro. doi: 10.1145/2488388.2488491. Cited on page 86.

Sofaer, H. R., Hoeting, J. A. & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. doi: 10.1111/2041-210X.13140. Cited on page 151.

Statistical Commission of the United Nations. (2014). *Fundamental Principles of Official Statistics.* `https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx`. (Accessed January 2021.) Cited on page 3.

Strichartz, R. S. (2000). *The way of analysis* (Revised Edition ed.). Jones and Bartlett. Cited on page 23.

Struijs, P., Braaksma, B. & Daas, P. J. H. (2014). Official statistics and big data. *Big Data & Society*, *1*(1), 1–6. doi: 10.1177/2053951714538417. Cited on page 2.

Taniguchi, M. & Tresp, V. (1997). Averaging regularized estimators. *Neural Computation*, *9*(5), 1163–1178. doi: 10.1162/neco.1997.9.5.1163. Cited on page 86.

Tarasconi, G. & Menon, C. (2017). *Matching Crunchbase with patent data* (OECD Science, Technology and Industry Working Papers No. 2017-07). Paris: OECD. Cited on page 107.

Ten Bosch, O. & Windmeijer, D. (2018, May). *Web scraping enterprise statistics. ESSNET Big Data Work Package 2 Deliverable 2.4 Final Report, pages 41–44.* `https://ec.europa.eu/eurostat/cros/sites/crosportal/files/Wp2_Del2_4.pdf`. (Accessed January 2021.) Cited on pages 113 and 122.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, *65*(331), 1350–1361. doi: 10.1080/01621459.1970.10481170. Cited on pages 10, 57, and 58.

Van Dantzig, D. (1951). On the consistency and the power of Wilcoxon's two sample test. *Proceedings of the KNAW, Series A*, *54*(1), 1–8. Cited on page 155.

Van Delden, A., Scholtus, S. & Burger, J. (2015). *Quantifying the effect of classification errors on the accuracy of mixed-source statistics* (Discussion Paper No. 201510). The Hague/Heerlen: Statistics Netherlands. Cited on pages 127 and 142.

Van Delden, A., Scholtus, S. & Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics*, *32*(3), 619–642. Cited on pages 8, 21, 60, 82, 84, 85, 87, 94, 109, 124, 125, 126, 141, and 185.

Van den Brakel, J. A. & Bethlehem, J. G. (2008). *Model-based estimation for official statistics* (Discussion Paper No. 08002). Voorburg/Heerlen: Statistics Netherlands. Cited on page 3.

Vanderlooy, S. & Hüllermeier, E. (2008). A critical analysis of variants of the AUC. *Machine Learning*, *72*(3), 247–262. doi: 10.1007/s10994-008-5070-x. Cited on pages 13, 152, 153, 156, and 157.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press. doi: 10.1017/CBO9780511802256. Cited on page 153.

Veran, S., Kleiner, K. J., Choquet, R., Collazo, J. A. & Nichols, J. D. (2012). Modeling habitat dynamics accounting for possible misclassification. *Landscape Ecology*, *27*(7), 943–956. doi: 10.1007/s10980-012-9746-z. Cited on page 86.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L. & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery*, *30*(4), 964–994. doi: 10.1007/s10618-015-0448-4. Cited on pages 4, 58, 69, and 185.

Widmer, G. & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*(1), 69–101. doi: 10.1023/A:1018046501280. Cited on page 58.

Wiedemann, G. (2019). Proportional classification revisited: Automatic content analysis of political manifestos using active learning. *Social Science Computer Review*, *37*(2), 135–159. doi: 10.1177/0894439318758389. Cited on page 19.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. doi: 10.2307/3001968. Cited on pages 151 and 154.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (p. 354-359). Cited on page 141.

Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. doi: 10.1016/C2015-0-02071-8. Cited on pages 120 and 122.

Yan, L., Dodier, R., Mozer, M. C. & Wolniewicz, R. (2003). Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In T. E. Fawcett & N. Mishra (Eds.), *Proceedings of the 20th International Conference on Machine Learning (ICML)* (pp. 848–855). Washington, D.C.. Cited on pages 13, 152, and 154.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In R. Greiner & D. Schuurmans (Eds.), *Proceedings of the 21st International Conference on Machine Learning (ICML)*. Banff. doi: 10.1145/1015330.1015425. Cited on page 86.

SUMMARY

In Chapter 1, two conflicting developments that affect the field of official statistics are identified. On the one hand, there is an increasing demand for the swift availability of detailed and highly accurate statistical information. The current craving for accurate information about excess deaths due to COVID-19 is a striking example. On the other hand, national statistical institutes (NSIs) that produce official statistics on such topics have to endure budget cuts and are obliged to reduce the survey burden on companies and citizens. The consequence of these two conflicting developments is that NSIs will have to rely increasingly on new types of data (i.e., big data) that must be processed and analysed by new types of methods (viz. statistical learning methods).

This thesis focuses on a specific group of statistical learning methods, namely classifiers. When the output of a classifier is aggregated, one obtains classifier-based statistics. If a classifier is not perfect, the resulting classifier-based statistics suffer from misclassification bias. To correct for that bias, a test set containing perfect information on the true classifications is required. A key challenge is selecting a correction method, in particular when dealing with time series that are non-stationary (i.e., that suffer from concept drift). In Chapter 1, the following open problem in the literature is raised: no solid theoretical analyses of methods correcting for misclassification bias in finite populations exist. Hence, the problem statement is formulated as follows.

> **Problem statement**: *In what way can we reduce misclassification bias in statistical learning so that we obtain more accurate classifier-based statistics?*

Next, two theoretical research questions and one empirical research question are derived from the problem statement. They are stated below and are complemented with the results that were obtained when addressing them.

Theoretical results. The simplest classifier-based statistic is the base rate of a dichotomous variable, i.e., the relative occurrence of a category among a total of two categories. In general, reducing bias increases variance. This phenomenon is referred to as the bias-variance trade-off. Therefore, the mean squared (estimation) error (MSE) is used to evaluate the accuracy of classifier-based statistics. The first research question is formulated as follows.

> **Research question 1**: *Which estimator of the base rate, in particular when dealing with concept drift, has the smallest MSE in finite populations?*

In Chapter 2, it is assumed that the test set that is used to correct for misclassification bias is a random sample from the population. Under that assumption, analytic expressions are derived for the MSE of two popular methods that correct for misclassification bias: (1) the misclassification estimator and (2) the calibration estimator. The expressions are valid up to and including terms of order $1/n$, where n is the sample size of the test set. It is shown that the MSE of the calibration estimator is always smaller than that of the misclassification estimator.

In Chapter 3, the main assumption from Chapter 2 is dropped. The assumption that the misclassification probabilities are identical in (a) the test set and (b) the unlabelled set is retained. This allows for prior probability shift, a specific type of concept drift. The theoretical derivations from Chapter 2 are adapted to the setting of Chapter 3. Next, numerical analyses are performed based on the derived analytic expressions. The main result is that the difference in MSE between the misclassification estimator and the calibration estimator is often in favour of the misclassification estimator when prior probability shift occurs.

The misclassification probabilities are estimated based on the test set. If the test set is small, the variance of the estimator is relatively large. A Bayesian framework is considered to investigate the variance of the estimator, leading to the formulation of the second research question.

> **Research question 2**: *How can we leverage identification regions of misclassification probabilities in order to reduce the MSE of classifier-based statistics even further?*

In Chapter 4 it is first demonstrated that the misclassification estimator might result in impermissible estimates of classifier-based statistics (e.g., negative counts) when small test sets are used. To prevent impermissible estimates, constraints on the prior distributions of misclassification probabilities are imposed. The constraints are based on the identification regions of the model parameters. Finally, a simulation study shows that imposing these parameter constraints reduces the MSE of the misclassification estimator even further.

Empirical results. Over the last decades, many statistical learning methods have been developed. The empirical evidence in fields other than official statistics is quite promising. However, the field of official statistics aims at accurate *aggregated* data, while many other fields aim at accurate predictions for *individual* data points. In Chapter 1 it was shown that these two types of accuracy are in some way each other's opposites. Therefore, empirical evidence of statistical learning methods in the field of official statistics is still required. Hence, a specific application in official statistics is considered, namely estimating cross-border Internet purchases within the European Union (EU). The third research question is formulated as follows.

> **Research question 3**: *To what extent can statistical learning be used to improve the accuracy of estimates of cross-border Internet purchases within the EU?*

Chapter 5 proposes a new methodology to estimate cross-border Internet purchases within the EU. The methodology is based on supply-side data, because demand-side data are argued to result in serious underestimations. Moreover, a combination of approximate string matching, locality-sensitive hashing, web scraping, and statistical learning is proposed as part of the new methodology. Subsequently, the methodology is applied to the Netherlands for the year 2016 leading to a rather surprising result: earlier estimates of cross-border Internet purchases within the EU (for the Netherlands in 2016) were less than a sixth of the actual value. These empirical results undeniably show that official statistics may be improved by implementing statistical learning methods.

Ranking instead of classifying. Empirical evidence from quantification learning (i.e., the median sweep method) shows that ranking objects instead of classifying objects might improve the accuracy of what we called classifier-based statistics. Investigating what we might call ranker-based statistics is left as a future research direction. Nonetheless, the problem of selecting the best ranker among a set of rankers is investigated in Chapter 6. In general, model selection requires a performance metric. The area under the receiver operating characteristic curve (AUC) is a performance metric that is typically used for model selection of rankers. In Chapter 6, preliminary theoretical evidence is provided showing that a smoothed variant of the AUC might be a better model selector of rankers than the standard AUC. That theoretical evidence seems to contradict earlier empirical evidence.

Conclusion. The conclusion of this thesis is that statistical learning methods can be used in the field of official statistics as long as misclassification bias is adequately corrected for. Our recommendation is to implement statistical learning methods (and the correction methods for misclassification bias discussed in this thesis) either to create new official statistics or to improve existing ones. Finally, we argue that domain experts are of vital importance to the successful implementation of statistical learning methods within official statistics.

SAMENVATTING

In Hoofdstuk 1 worden twee tegenstrijdige ontwikkelingen aangewezen die effect hebben op officiële statistiek. Aan de ene kant is er een toenemende behoefte aan de snelle beschikbaarheid van gedetailleerde en betrouwbare statistische informatie. De huidige honger naar betrouwbare informatie omtrent de oversterfte ten gevolge van COVID-19 is daarvan een treffend voorbeeld. Aan de andere kant hebben officiële statistiekbureaus die dergelijke officiële statistieken produceren (zoals het CBS) te maken met bezuinigingen en de verplichting om administratieve lasten te verlagen. Het gevolg van deze twee tegenstrijdige ontwikkelingen is dat statistiekbureaus in toenemende mate afhankelijk zijn van nieuwe soorten data (denk aan big data) die alleen verwerkt en geanalyseerd kunnen worden met behulp van nieuwe soorten methoden (waaronder statistical learning methods).

Dit proefschrift richt zich op een specifieke groep van statistical learning methods, namelijk classifiers. De geaggregeerde uitkomsten van een classifier noemen wij classifier-based statistics. Als de gebruikte classifier niet foutloos is, dan treedt er misclassification bias op. Om voor die vertekening te kunnen corrigeren is een test set nodig waarin foutloze informatie staat over de te voorspellen klassen. Het is vervolgens een grote uitdaging om een juiste correctiemethode te kiezen. Dat geldt in het bijzonder voor tijdreeksanalyse waarbij de data niet stationair zijn (of, met andere woorden, leidt onder concept drift). In Hoofdstuk 1 wordt het volgende open probleem in de literatuur boven water gehaald: er bestaat voor eindige populaties geen gedegen theoretische analyse van methodes die corrigeren voor misclassification bias. Hieruit volgt de onze probleemstelling.

> **Probleemstelling**: *Op welke manier kunnen we misclassification bias in statistical learning verminderen opdat we classifier-based statistics met hogere nauwkeurigheid verkrijgen?*

Vervolgens worden er twee theoretische en een empirische onderzoeksvraag uit deze probleemstelling afgeleid. Ze worden hieronder genoemd en aangevuld met de resultaten die behaald zijn tijdens het behandelen van de onderzoeksvragen.

Theoretische resultaten. De meest eenvoudige classifier-based statistics is de base rate van een dichotome variabele, dat wil zeggen, de relatieve frequentie van één categorie in een groep van twee categorieën. In het algemeen zorgt het verlagen van bias voor het verhogen van variantie. Dit verschijnsel wordt de bias-variance trade-off genoemd. Daarom gebruiken wij de mean squared (estimation) error (MSE) om de nauwkeurigheid van classifier-based statistics te toetsen. De eerste onderzoeksvraag is als volgt geformuleerd.

> **Onderzoeksvraag 1**: *Welke schatter voor de base rate heeft de laagste MSE in eindige populaties, in het bijzonder als we te maken hebben met concept drift.*

In Hoofdstuk 2 wordt aangenomen dat de test set (die we gebruiken om te corrigeren voor misclassification bias) een aselecte steekproef is uit de populatie. Onder die aanname worden analytische uitdrukkingen afgeleid voor de MSE van twee populaire correctietechnieken voor misclassification bias: (1) de misclassification estimator en (2) de calibration estimator. De uitdrukkingen zijn benaderingen die geldig zijn voor alle termen tot die van order $1/n$, waarbij n de omvang van de test set aanduidt. Het wordt aangetoond dat de MSE van de calibration estimator altijd kleiner is dan die van de misclassification estimator.

In Hoofdstuk 3 wordt een belangrijke aanname uit Hoofdstuk 2 losgelaten. Wel behouden we de aannames dat de misclassification probabilities gelijk zijn in zowel (a) de test set als (b) de unlabelled set. Hierdoor wordt prior probability shift, een specifieke vorm van concept drift, weer enigszins toegelaten. De theoretische afleidingen uit Hoofdstuk 2 worden aangepast naar de setting van Hoofdstuk 3. Vervolgens worden er numerieke analyses uitgevoerd met de verkregen analytische uitdrukkingen als basis. Het hoofdresultaat van Hoofdstuk 3 is dat het verschil in MSE tussen de misclassification estimator en calibration estimator vaak in het voordeel is van de misclassification estimator, met name in het geval van prior probability shift.

De misclassification probabilities worden geschat op basis van een test set. Als de test set klein is, dan is de variantie van die schatter relatief groot. We beschouwen een Bayesiaans raamwerk om de variantie van die schatter te onderzoeken. Dit heeft geleid tot de tweede onderzoeksvraag.

> **Onderzoeksvraag 2**: *Hoe kunnen we identification regions benutten om misclassification bias te verminderen.*

In Hoofdstuk 4 wordt getoond dat de misclassification estimator tot ontoelaatbare schattingen van classifier-based statistics (zoals negatieve tellingen) kan leiden. Om zulke schattingen te voorkomen wordt voorgesteld om de parameter constraints te baseren op identification regions. Tot slot laat een simulatiestudie zien dat de parameter constraints de MSE van de misclassification estimator nog verder kunnen verlagen.

Empirische resultaten. In de laatste jaren zijn veel statistical learning methoden ontwikkeld. De empirische evidentie in vakgebieden buiten de officiële statistiek is veelbelovend. Echter, binnen de officiële statistiek zijn we gericht op nauwkeurige data op *geaggregeerd* niveau, terwijl veel andere vakgebieden geïnteresseerd zijn in nauwkeurige voorspellingen voor *individuele* datapunten. In Hoofdstuk 1 werd al aangetoond dat deze twee typen nauwkeurigheid in zekere zin elkaars tegenovergestelde zijn. De empirische evidentie van statistical learning methoden binnen de officiële statistiek staat dan ook nog in de kinderschoenen. Met dat in gedachte bestuderen wij een specifieke toepassing binnen de officiële statistiek, namelijk het schatten van grensoverschrijdende internetaankopen binnen de Europese Unie (EU). De derde onderzoeksvraag luidt als volgt.

> **Onderzoeksvraag 3**: *In hoeverre kan statistical learning gebruikt worden om de nauwkeurigheid van grensoverschrijdende internetaankopen binnen de EU te verbeteren?*

Hoofdstuk 5 stelt een nieuwe methodologie voor om grensoverschrijdende internetaankopen binnen de EU te schatten. De methodologie is gebaseerd op gegevens vanuit de productiekant, omdat data vanuit de consumptiekant leiden tot zeer grote onderschattingen. Bovendien wordt als onderdeel van de methodologie een combinatie van approximate string matching, locality-sensitive hashing, web scraping, and statistical learning voorgesteld. Vervolgens wordt de methodologie toegepast voor Nederland in het jaar 2016 met een verrassend resultaat: eerdere schattingen van grensoverschrijdende internetaankopen binnen de EU blijken minder dan een zesde van de werkelijke waarde te vertegenwoordigen. Dit empirische bewijs laat ontegenzeggelijk zien dat officiële statistieken verbeterd kunnen worden door statistical learning methoden te implementeren.

Rangschikken boven classificeren. Empirisch bewijs vanuit quantification learning (zoals de median sweep methode) laat ziet dat het rangschikken (Engels: ranking) van objecten, afgezet tegen het classificeren van objecten, de nauwkeurigheid van wat wij classifier-based statistics noemen kan verbeteren. Het onderzoek naar wat je dan ranker-based statistics zou kunnen noemen laten we over aan toekomstig onderzoek. We kijken in Hoofdstuk 6 echter wel naar het probleem van het kiezen van een goede ranker binnen een gegeven groep rankers. In het algemeen vereist dergelijke modelselectie om een performance metric. Bij het vergelijken van rankers wordt doorgaans de zogenaamde area under the receiver operating characteristic curve (AUC) gebruikt. In Hoofdstuk 6 presenteren we voorlopige theoretische resultaten die laten zien dat gladde (Engels: smoothed) varianten van de AUC niet onder doen voor de standaard AUC. Die theoretische resultaten lijken eerdere empirische resultaten uit de literatuur tegen te spreken.

Conclusie. De conclusie van dit proefschrift is dat statistical learning methoden zeker gebruikt kunnen worden voor officiële statistiek, zolang er maar op de juiste wijze voor misclassification bias wordt gecorrigeerd. Onze aanbeveling is om statistical learning methoden vooral in te zetten om nieuwe of verbeterde officiële statistieken te produceren (gebruikmakend van de correctiemethoden voor misclassification bias zoals besproken in dit proefschrift). Ten slotte betogen we dat experts met domeinkennis onmisbaar zijn voor het succesvol inzetten van statistical learning methoden binnen de officiële statistiek.

CURRICULUM VITAE

Quinten Alexander Meertens was born in Amsterdam, the Netherlands, on 16 December 1993. He graduated cum laude from OSG De Meergronden in Almere in 2010. Later that year, he started his studies in theoretical mathematics at the University of Amsterdam. He obtained his MSc degree (cum laude) in 2015.

With a strong desire to work on social issues, he started as a statistical researcher at Statistics Netherlands after the summer of 2015. He worked on different data science projects, including the estimation of cross-border Internet purchases within the EU. These efforts resulted in receiving the Statistics Netherlands Division of Economic Statistics' Innovation Award in 2016.

In the spring of 2016, he decided to aim for a PhD position within Statistics Netherlands. With funding from both the Division of Economic Statistics and the Research Department of Statistics Netherlands, he started his PhD in October 2016 under the supervision of Cees Diks, Jaap van den Herik and Frank Takes. The main goal of the PhD research was to enrich official economic statistics using data-driven modelling techniques from both econometrics and computer science.

During his PhD, Quinten has developed and implemented a new methodology to estimate cross-border Internet purchases within the EU. The results uncovered some serious biases in existing estimates. As a consequence, he has been invited to present his findings to three international working groups: (1) the OECD Working Party on Measurement and Analysis of the Digital Economy, (2) the OECD Informal Advisory Group on Measuring GDP in a Digitalised Economy and (3) the ECB Working Group on General Economic Statistics.

In the last year of his PhD, he worked at the Research Department of Statistics Netherlands focusing on more fundamental research questions. Together with Sander Scholtus and Julian Karch he supervised Kevin Kloos, leading to a nomination for the Best Paper Award at BNAIC/BENELEARN 2020.

As of November 2020, he is head of Justice and Safety Statistics.

PUBLICATIONS

The content of this thesis is based on the following four publications. The next page shows publications that resulted from other collaborations during my PhD.

- Meertens, Q. A., Diks, C. G. H., Van den Herik, H. J. & Takes, F. W. (2020). Improving the output quality of official statistics based on machine learning algorithms. Submitted to *Journal of Official Statistics*.
- Kloos, K., Meertens, Q. A., Scholtus, S. & Karch, J. D. (2020). Comparing correction methods to reduce misclassification bias. In L. Cao, W. A. Kosters & J. Lijffijt (Eds.), *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC)* (pp. 103-129). Leiden.
    - *My contribution: deriving the proof of the fourth theorem (see Theorem 2.4), reviewing the literature together with Kevin, and writing most of the paper.*
- Meertens, Q. A., Diks, C. G. H., Van den Herik, H. J. & Takes, F. W. (2020). 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.
- Meertens, Q. A., Diks, C. G. H., Van den Herik, H. J. & Takes, F. W. (2019). A Bayesian approach for accurate classification-based aggregates. In T. Y. Berger-Wolf & N. V. Chawla (Eds.), *Proceedings of the 19th SIAM International Conference on Data Mining (SDM)* (pp. 306–314). Calgary.

During my PhD, I have had the pleasure to collaborate with other colleagues and scientists on different projects. The following publications are the result of these projects.

- Oostenbroek, M. H. J., Van der Leij, M. J., Meertens, Q. A., Diks, C. G. H. & Wortelboer, H. M. (2021). Link-based influence maximization in networks of health promotion professionals. Submitted to *PLoS ONE*.
  - *My contribution: deriving the theoretical results and co-writing the paper.*
- Delden, A., Scholtus, S., Burger, J. & Meertens, Q. A. (2020). Accuracy of estimated ratios as affected by dynamic classification errors. Submitted to *Journal of the American Statistical Association*.
  - *My contribution: reviewing the literature.*
- Burger, J. & Meertens, Q. A. (2020). The algorithm versus the chimps: On the minima of classifier performance metrics. In L. Cao, W. A. Kosters & J. Lijffijt (Eds.), *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC)* (pp. 38-55). Leiden.
  - *My contribution: deriving the theoretical results and co-writing the paper.*

DANKWOORD

Dit proefschrift was er niet geweest zonder de hulp en steun van een groot aantal mensen. Als eerste wil ik graag mijn begeleiders Cees, Jaap en Frank bedanken voor een bijzonder leerzame en leuke tijd. Cees, ik vond het altijd erg gezellig om op het instituut bij te praten of als het mooi weer was naast Artis te lunchen. Je hebt me geleerd niet bang te hoeven zijn voor econometristen en mij altijd gestimuleerd en geprikkeld om kritisch naar mijn werk te kijken. Jaap, wij hadden als wiskundigen en perfectionisten naar mijn idee al direct een klik. De wijze waarop jij naar woorden en zinnen kijkt heeft mij onvoorstelbaar veel geleerd over het schrijven van een goede tekst. Ik wil je oprecht bedanken voor je behulpzaamheid en jouw drive om het beste in mij naar boven te halen. Frank, ik heb me dankzij jou altijd welkom gevoeld in Leiden. Onze gesprekken waren gezellig maar ook bijzonder efficiënt. Je hebt me veel geleerd over het onderscheiden van hoofdzaken en bijzaken en daar ben ik je zeer dankvoor.

Tijdens mijn promotieonderzoek heb ik het genoegen gehad om met een aantal mensen nauw samen te werken. Arjan, we hebben samen toch maar mooi een nieuwe statistiek opgezet. Ik bewonder jouw doortastendheid en bedank je graag voor alles wat je me hebt geleerd over statistiekproductie. Willem, graag bedank ik je voor je flexibiliteit en doorzettingsvermogen, je was onmisbaar. Joep, het schrijven van ons artikel over *the chimps* vond ik erg leuk, bedankt dat ik mee mocht doen. Arnout, bedankt voor al je feedback en voor je enorme drive. Ik kijk terug op leerzame samenwerkingen en veel inspirerende gesprekken. Sander, bedankt voor de nieuwe bewijstechnieken die je me hebt geleerd en voor de prettige co-begeleiding van Kevin. Dan jij Kevin, bedankt voor je zeer harde werk en voor de gezelligheid. Marco, Heleen en Maurits, ontzettend leuk om met jullie en Cees een artikel te schrijven over netwerken in health promotion programs. Bedankt dat ik hier een rol in mocht spelen.

Daarnaast heb ik met veel plezier gewerkt in de omgeving van leuke collega's. Om te beginnen natuurlijk de collega's van team B&C; heerlijk die goede humor in de kamer en op de gang. Dan team Methodologie Den Haag, bedankt voor het warme welkom en alle leerzame discussies. Ook bij CeNDEF en LIACS heb ik het naar mijn zin gehad. Hao, thank you for many nice and inspiring conversations. Florian, ik vond het erg prettig om met je samen te werken en ik waardeer het nog steeds dat je me de kans hebt gegeven een hoorcollege te geven. Gerrit-Jan, António, Anne, Daniela, het was altijd gezellig op de hoekkamer.

Drie mensen hebben nog onmisbare input voor mijn werk geleverd. Javier, thank you for guiding me through the ORBIS database and for posing the idea of hashing and combining string matching methods for the purpose of probabilistic record linkage. Holger, thank you for taking the time to provide feedback on my Bayesian work in machine learning. Eyke, vielen Dank, dass du mich auf die Quantification Learning Methoden hingewiesen hast. Dies hatte einen großen Einfluss auf die letzte Phase meiner Untersuchung.

Dan wil ik graag een aantal mensen bedanken zonder wie ik dit traject nooit had kunnen starten. Danny, dankzij jou is dit allemaal begonnen. Je hebt altijd het volle vertrouwen in mij gehad en me door weer en wind ondersteund in dit traject. Ilanah, je hebt je handen voor mij in het vuur gestoken en mijn promotieplek financieel mogelijk gemaakt. Daarvoor ben ik je nog altijd zeer dankbaar. Barteld, bedankt voor het contact leggen tussen mij en mijn begeleiders en voor alle inspirerende gesprekken. Jacobiene, graag bedank ik jou voor je enorme inzet tijdens dit traject, zonder jou had ik het nooit georganiseerd gekregen.

Tot slot richt ik mij tot mijn vrienden en familie. Marc, Sho, Laurens, Miriam, Floris, Daniël, Sebastiaan, bedankt voor jullie vriendschap en voor alle ontspannende momenten. Lieve VV'ers, hopelijk kunnen we elkaar nog lang blijven inspireren. Lieve wiskundigen, bedankt voor de leuke studententijd en sindsdien de gezellige kerstdiners. Dan wil ik nu graag mijn vader bedanken. Papa, ik vind het zo jammer dat je dit niet hebt mogen meemaken en ik mis je enorm. Je hebt me altijd gestimuleerd om mezelf te zijn en me van mijn beste kant te laten zien. Het is voor mij heel speciaal om op jouw 65e geboortedag te verdedigen. Tycho, Oscar en Merel, bedankt voor jullie steun de afgelopen jaren, voor de fijne en soms moeilijke gesprekken, maar ook voor alle gezellige spelletjesavonden en etentjes. Ankie en Reinier, Sanne en Tom, en natuurlijk Pien en Mats, bedankt dat jullie me als je eigen familie hebben omarmd. Nils, Olivier en Thijmen, wat ben ik trots op jullie. Jullie zijn mijn alles. Myrte, je bent mijn enige echte. Bedankt voor je onvoorwaardelijke liefde en steun. Ik ben je dankbaar voor je geduld, maar je zet me ook op scherp. Zonder jou was dit proefschrift er nooit gekomen.

SIKS DISSERTATION SERIES

2011 01 Botond Cseke (RUN), Variational Algorithms for Bayesian Inference in Latent Gaussian Models

02 Nick Tinnemeier (UU), Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

03 Jan Martijn van der Werf (TUE), Compositional Design and Verification of Component-Based Information Systems

04 Hado van Hasselt (UU), Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference

05 Bas van der Raadt (VU), Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.

06 Yiwen Wang (TUE), Semantically-Enhanced Recommendations in Cultural Heritage

07 Yujia Cao (UT), Multimodal Information Presentation for High Load Human Computer Interaction

08 Nieske Vergunst (UU), BDI-based Generation of Robust Task-Oriented Dialogues

09 Tim de Jong (OU), Contextualised Mobile Media for Learning

10 Bart Bogaert (UvT), Cloud Content Contention

11 Dhaval Vyas (UT), Designing for Awareness: An Experience-focused HCI Perspective

12 Carmen Bratosin (TUE), Grid Architecture for Distributed Process Mining

13 Xiaoyu Mao (UvT), Airport under Control. Multiagent Scheduling for Airport Ground Handling

14 Milan Lovric (EUR), Behavioral Finance and Agent-Based Artificial Markets

15 Marijn Koolen (UvA), The Meaning of Structure: the Value of Link Evidence for Information Retrieval

16 Maarten Schadd (UM), Selective Search in Games of Different Complexity

17 Jiyin He (UVA), Exploring Topic Structure: Coherence, Diversity and Relatedness

18 Mark Ponsen (UM), Strategic Decision-Making in complex games

19 Ellen Rusman (OU), The Mind's Eye on Personal Profiles

20 Qing Gu (VU), Guiding service-oriented software engineering - A view-based approach

21 Linda Terlouw (TUD), Modularization and Specification of Service-Oriented Systems

22 Junte Zhang (UVA), System Evaluation of Archival Description and Access

23 Wouter Weerkamp (UVA), Finding People and their Utterances in Social Media

24 Herwin van Welbergen (UT), Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

25 Syed Waqar ul Qounain Jaffry (VU), Analysis and Validation of Models for Trust Dynamics

26 Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

27 Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

28 Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

29 Faisal Kamiran (TUE), Discrimination-aware Classification

30 Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

31 Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

32 Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

33 Tom van der Weide (UU), Arguing to Motivate Decisions

34 Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

35 Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

36 Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

37 Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

38 Nyree Lemmens (UM), Bee-inspired Distributed Optimization

39 Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

40 Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

41 Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

42 Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

43 Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

44 Boris Reuderink (UT), Robust Brain-Computer Interfaces

45 Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

46 Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

47 Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

48 Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

49 Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012

01 Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

02 Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

03 Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

04 Jurriaan Souer (UU), Development of Content Management System-based Web Applications

05 Marijn Plomp (UU), Maturing Interorganisational Information Systems

06 Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

07 Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

08 Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms

10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment

11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems

13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems

15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.

16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance

18 Eltjo Poort (VU), Improving Solution Architecting Practices

19 Helen Schonenberg (TUE), What's Next? Operational Support for Business Process Execution

20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval

22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval

25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

26 Emile de Maat (UVA), Making Sense of Legal Text

27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games

28 Nancy Pascall (UvT), Engendering Technology Empowering Women

29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval

30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making

31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure

32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning

33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)

34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications

35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics

36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes

37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation

38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms

39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks

40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia

41 Sebastian Kelle (OU), Game Design Patterns for Learning

42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning

43 Withdrawn

44 Anna Tordai (VU), On Combining Alignment Techniques

45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions

46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation

47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior

48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data

49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions

50 Steven van Kervel (TUD), Ontologogy driven Enterprise Information Systems Engineering

51 Jeroen de Jong (TUD), Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

---

2013 01 Viorel Milea (EUR), News Analytics for Financial Decision Support

02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing

03 Szymon Klarman (VU), Reasoning with Contexts in Description Logics

04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling

05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns

06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience

07 Giel van Lankveld (UvT), Quantifying Individual Player Differences

08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators

09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications

10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.

11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services

12 Marian Razavian (VU), Knowledge-driven Migration to Services

13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly

14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning

15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications

16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation

17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid

18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification

19 Renze Steenhuizen (TUD), Coordinated Multi-Agent Planning and Scheduling

20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval

21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation

22 Tom Claassen (RUN), Causal Discovery and Logic

23 Patricio de Alencar Silva (UvT), Value Activity Monitoring

24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning

25 Agnieszka Anna Latoszek-Berendsen (UM), Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System

26 Alireza Zarghami (UT), Architectural Support for Dynamic Homecare Service Provisioning

27 Mohammad Huq (UT), Inference-based Framework Managing Data Provenance

28 Frans van der Sluis (UT), When Complexity becomes Interesting: An Inquiry into the Information eXperience

29 Iwan de Kok (UT), Listening Heads

30 Joyce Nakatumba (TUE), Resource-Aware Business Process Management: Analysis and Support

31 Dinh Khoa Nguyen (UvT), Blueprint Model and Language for Engineering Cloud Applications

32 Kamakshi Rajagopal (OUN), Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

33 Qi Gao (TUD), User Modeling and Personalization in the Microblogging Sphere

34 Kien Tjin-Kam-Jet (UT), Distributed Deep Web Search

35 Abdallah El Ali (UvA), Minimal Mobile Human Computer Interaction

36 Than Lam Hoang (TUe), Pattern Mining in Data Streams

37 Dirk Börner (OUN), Ambient Learning Displays

38 Eelco den Heijer (VU), Autonomous Evolutionary Art

39 Joop de Jong (TUD), A Method for Enterprise Ontology based Design of Enterprise Information Systems

40 Pim Nijssen (UM), Monte-Carlo Tree Search for Multi-Player Games

41 Jochem Liem (UVA), Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

42 Léon Planken (TUD), Algorithms for Simple Temporal Reasoning

43 Marc Bron (UVA), Exploration and Contextualization through Interaction and Concepts

---

2014 01 Nicola Barile (UU), Studies in Learning Monotone Models from Data

02 Fiona Tuliyano (RUN), Combining System Dynamics with a Domain Modeling Method

03 Sergio Raul Duarte Torres (UT), Information Retrieval for Children: Search Behavior and Solutions

04 Hanna Jochmann-Mannak (UT), Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

05 Jurriaan van Reijsen (UU), Knowledge Perspectives on Advancing Dynamic Capability

06 Damian Tamburri (VU), Supporting Networked Software Development

07 Arya Adriansyah (TUE), Aligning Observed and Modeled Behavior

08 Samur Araujo (TUD), Data Integration over Distributed and Heterogeneous Data Endpoints

09 Philip Jackson (UvT), Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

10 Ivan Salvador Razo Zapata (VU), Service Value Networks

11 Janneke van der Zwaan (TUD), An Empathic Virtual Buddy for Social Support

12 Willem van Willigen (VU), Look Ma, No Hands: Aspects of Autonomous Vehicle Control

13 Arlette van Wissen (VU), Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

14 Yangyang Shi (TUD), Language Models With Meta-information

15 Natalya Mogles (VU), Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria

17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations

19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support

20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link

21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments

22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training

23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era

24 Davide Ceolin (VU), Trusting Semi-structured Web Data

25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction

26 Tim Baarslag (TUD), What to Bid and When to Stop

27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty

28 Anna Chmielowiec (VU), Decentralized k-Clique Matching

29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software

30 Peter de Cock (UvT), Anticipating Criminal Behaviour

31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support

32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data

33 Tesfa Tegegne (RUN), Service Discovery in eHealth

34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.

35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach

36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models

37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying

38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.

39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital

40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education

41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text

42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models

43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments

44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.

45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach

46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity

47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015 01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response

02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
04 Howard Spoelstra (OUN), Collaborations in Open Learning Environments
05 Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding
06 Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
07 Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis
08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
09 Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
10 Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning
11 Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins
12 Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks
13 Giuseppe Procaccianti (VU), Energy-Efficient Software
14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
15 Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation
16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
19 Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners
20 Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination
21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
22 Zhemin Zhu (UT), Co-occurrence Rate Networks
23 Luit Gazendam (VU), Cataloguer Support in Cultural Heritage
24 Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation
25 Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection
26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
27 Sándor Héman (CWI), Updating compressed colomn stores
28 Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO
29 Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
30 Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning
31 Yakup Koç (TUD), On the robustness of Power Grids
32 Jerome Gard (UL), Corporate Venture Management in SMEs
33 Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources
34 Victor de Graaf (UT), Gesocial Recommender Systems

35 Jungxao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

---

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UVA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web

19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UVA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UVA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (UvT), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UVA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TUE), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UVA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VU), Logics for causal inference under uncertainty

23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (UvT), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TUE), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TUE), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech

30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data

06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VU), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming

36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TUE), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst

31 Gongjin Lan (VU), Learning better – From Baby to Better

32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
	02	Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
	03	Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
	04	Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
	05	Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
	06	Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
	07	Armel Lefebvre (UU), Research data management for open science
	08	Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
	09	Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play