

Statistical Advances in Data Linkage and Model Evaluation

by

Olivier Binette

Department of Statistical Science
Duke University

Defense Date: July 3, 2024

Approved:

Jerome P. Reiter, Supervisor

David Banks

Eric B. Laber

Sean Michael O'Brien

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2024

ABSTRACT

Statistical Advances in Data Linkage and Model Evaluation

by

Olivier Binette

Department of Statistical Science
Duke University

Defense Date: July 3, 2024

Approved:

Jerome P. Reiter, Supervisor

David Banks

Eric B. Laber

Sean Michael O'Brien

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2024

Copyright © 2024 by
Olivier Binette
All rights reserved

Abstract

This dissertation is about statistical contributions to data linkage and model evaluation. The two subjects fall at the extremities of traditional model development, with data linkage used to enrich data fed into downstream models and analyses, and evaluation used to maximize the utility of deployed models. We report on five research projects where we developed generalizable statistical methodologies to solve important practical problems in these areas. This includes the evaluation of statistical models for the quantification of modern slavery, methods to estimate and monitor the generalization performance of entity resolution systems, a novel F-score optimization algorithm for bipartite record linkage, and the introduction of an estimands framework to improve the validity and practical usefulness of AI/ML evaluations.

Contents

Abstract	iv
List of Tables	x
List of Figures	xii
Acknowledgements	xiv
1 Introduction	1
1.1 Motivation	1
1.2 Structure and Overview	2
2 On the Reliability of Multiple Systems Estimation for the Quantification of Modern Slavery	4
2.1 Introduction	5
2.1.1 Two Centuries of Controversy	6
2.1.2 Our Contribution	7
2.1.3 Organization of the Paper	8
2.2 Background	8
2.2.1 Introduction to Multiple Systems Estimation	9
2.2.2 The General Framework of Multiple Systems Estimation	10
2.2.3 Assumptions and Consistency of Estimators	13
2.2.4 Multiple Systems Estimation Methods	16
2.3 Data and Population Size Estimators Under Consideration	22
2.3.1 Data From Past Modern Slavery Studies	22

2.3.2	Population Size Estimators	23
2.3.3	Comparison of Estimates	26
2.3.4	Sensitivity and Convergence Issues	27
2.4	Internal Consistency Analysis	33
2.4.1	Ground Truth Data Through Conditioning	33
2.4.2	Analysis and Results	33
2.4.3	Limitations	36
2.5	Bias Under Misspecified Assumptions	36
2.5.1	Characterization of the Asymptotic Relative Bias	37
2.5.2	Bias in the Presence of Individual Heterogeneity	39
2.5.3	Summary	42
2.6	Visual Assessment of Robustness	43
2.6.1	Visualizing Estimate Trajectories	43
2.6.2	Application to Real Data	45
2.7	Discussion	48
2.8	Appendix	50
2.8.1	Datasets Summary	50
2.8.2	Proof of Theorem 1	52
2.8.3	Proof of Proposition 1	56
3	Estimating the Performance of Entity Resolution Algorithms: Lessons Learned Through PatentsView.org	57
3.1	Introduction	57
3.1.1	The Evaluation Problem	60
3.1.2	Structure of the Paper	66
3.2	Data and Methodology	66
3.2.1	Benchmark Datasets for Inventor Disambiguation	66

3.2.2	Hand-Disambiguation Methodology	67
3.2.3	Proposed Performance Estimators	68
3.2.4	Simulation Study	73
3.3	Results	75
3.3.1	Results From the Simulation Study	75
3.3.2	Evaluation of PatentsView’s Disambiguation	81
3.4	Discussion	82
3.5	Appendix	84
3.5.1	Bias of Precision Computed on Benchmark Datasets	84
3.5.2	Precision and Recall Estimator Formulas	85
4	How to Evaluate Entity Resolution Systems: An Entity-Centric Framework with Application to Inventor Name Disambiguation	87
4.1	Introduction	87
4.1.1	Previous Work	90
4.1.2	Outline of the Paper	92
4.2	Background	92
4.2.1	Patent Data Disambiguation for PatentsView.org (♠)	92
4.2.2	Motivating Data and Application	93
4.2.3	Industry Standards for Entity Resolution Evaluation (♠)	96
4.3	Methodology	98
4.3.1	Summary Statistics and Quality Assurance	98
4.3.2	Data Labeling Methodology	102
4.3.3	Error Analysis	105
4.3.4	Performance Metric Estimation	110
4.4	Empirical Illustrations and Simulations	115
4.4.1	Summary Statistics and Quality Assurance	115

4.4.2	Performance Estimates	118
4.4.3	Error analysis	120
4.4.4	Simulation Study	123
4.5	Discussion	128
4.6	Appendix	130
5	Optimal F-score Clustering for Bipartite Record Linkage	134
5.1	Introduction	134
5.2	F -score Optimization Under a Bipartite Record Linkage Constraint	137
5.2.1	F -score Objective Function and Estimators	138
5.2.2	Algorithm for Approximating the F -score	140
5.3	Estimation of Overlap Size (\spadesuit)	144
5.3.1	Definition of the BRL Estimator (\spadesuit)	145
5.3.2	Conservative Nature of the BRL Estimator (\spadesuit)	147
5.4	Simulation Studies and Illustrative Examples	150
5.4.1	Bayesian Bipartite Record Linkage Model (\spadesuit)	151
5.4.2	Simulation Study (\spadesuit)	152
5.4.3	Illustrative Examples	153
5.5	Discussion	157
6	Improving the Validity and Practical Usefulness of AI/ML Evaluations Using an Estimands Framework	159
6.1	Introduction	160
6.2	Background	164
6.2.1	Definitions and Related Work	164
6.2.2	ML Evaluation as a Discipline	166
6.3	Three Examples Leading to Rank Reversals	166
6.3.1	Defining Performance Rank Reversals	167

6.3.2	Rank Reversals With Cross-Validation	168
6.3.3	Rank Reversals in Clustering Evaluation	170
6.3.4	Rank Reversals in LLM Benchmarking	173
6.4	Our Proposal: Better-Defined Targets of Estimation Through the Estimands Framework	176
6.4.1	The Estimands Framework	176
6.4.2	Application of the Estimands Framework to Rank Reversal Examples	178
6.5	Discussion	187
7	Conclusion	189
	Bibliography	190

List of Tables

2.1	Counts of potential victims of modern slavery in the UK, disaggregated by the lists in which potential victims appear.	12
2.2	Datasets under consideration.	23
2.3	Convergence diagnostics for LCMCR samples aggregated across chains.	32
2.4	United Kingdom dataset conditioned on the LA list.	34
2.5	Description of conditioned datasets with more than 30 observations. .	34
2.6	Summary results of the internal consistency analysis.	35
3.1	Bias and root mean squared error (rmse) of precision estimators. . . .	77
3.2	Bias and root mean squared error of recall estimators.	80
3.3	Estimated pairwise precision and recall.	81
4.1	Example of attributes available for individual inventor mentions. . . .	95
5.1	Example of a linkage scenario.	148
5.2	Example of a general scenario where the LSAP algorithm does not declare a link.	148
5.3	The \mathbf{m} and \mathbf{u} parameters for the three simulation scenarios.	152
5.4	Average F -score, estimated file overlap, true overlap, and model-based overlap 95% credible interval.	153
5.5	F -score, estimated file overlap, true overlap, and model-based overlap 95% credible interval.	155
5.6	F -score, estimated file overlap, true overlap, and model-based overlap 95% credible interval.	157
6.1	Summary of our cross-validation experiment.	170

6.2	Summary of the clustering evaluation experiment.	172
6.3	Example of a rank reversal between two models on the Format- Following benchmark dataset.	175

List of Figures

2.1	Examples of independence graphs on the 5 lists considered in the UK study of Silverman (2014); Bales et al. (2015).	20
2.2	Comparison of the independence model, SparseMSE, LCMCR and dga estimates.	27
2.3	Estimates of the SparseMSE approach on the United Kingdom data. .	29
2.4	Point estimates and 95% credible intervals of dga estimates for the United Kingdom dataset.	30
2.5	MCMC traces of the non-observation probability.	32
2.6	Results of the internal consistency analysis for every considered dataset and reference list.	35
2.7	Asymptotic relative bias of population size estimators.	42
2.8	Trajectories of dga estimates.	45
2.9	Visualization of estimate trajectories for the United Kingdom dataset.	46
2.10	Visualization of estimate trajectories for the Netherlands dataset. . .	47
3.1	Distribution of precision estimates versus the true precision.	62
3.2	Example of ground truth clustering and predicted clustering.	69
3.3	Distribution of precision estimates.	76
3.4	Distribution of recall estimates for various sample sizes and misattribution rates.	79
4.1	Diagram representation of the main elements of the framework and their dependencies.	91
4.2	Estimated number of citations to PatentsView.	94

4.3	Screenshot of the Streamlit app used for clerical error review.	109
4.4	Summary statistics for PatentsView’s history of predicted disambiguations.	116
4.5	Summary statistics and estimates for the fixed data set of inventor mentions dating up to August 2017.	118
4.6	Performance metrics estimates and confidence intervals over PatentsView’s disambiguation history.	119
4.7	Reviewer’s notes and their weighted relative frequencies for patterns in overclustering and underclustering errors.	121
4.8	Performance difference from the baseline for inventors with an inferred Asian and Pacific Islander ethnicity (API) or other inferred ethnicity.	122
4.9	Simulation study based on the RLdata10000 data set.	124
4.10	Distribution of the standard deviation estimator for pairwise precision and pairwise recall estimators.	127
4.11	Simulation study based on the PatentsView’s historical disambiguations.	133
5.1	Example of an augmented matrix $\tilde{\Delta}^{(2)}$ used in the LSAP.	142
6.1	Estimands framework adapted from ICH (2019) for ML model evaluation.	162
6.2	Comparing the distribution of features between the full California Housing Dataset and a training dataset.	180

Acknowledgements

I am privileged for the support I have received throughout my PhD, the connections I have made, and the experiences I have gained. My wife, family and friends will recognize themselves in the support and encouragement they have given me. I am also grateful for the support I received from Jerry Reiter, David Banks, Beka Steorts, Serge Aleshin-Guendel, Ted Enamorado, Bernard Silverman, Abel Dasylva, Giri Gopalan, Emily Casleton, Nidhi Parikh, Debdeep Pati, Merlise Clyde, David Dunson, Simon Guillotte, Jean-François Coeurjolly, Olivier Colin, and many others. My co-authors, including some of the above as well as Eric Bai, Youngsoo Baek, Sarvo Madhavan, Christina Jones, Emma Hickerson, Aida York, and Jack Butler, have made writing papers an enjoyable team experience. I am eternally grateful for the administrative staff in our department, Lori Raunch, Nicole Scott, and Karen Whitesell, that have always been readily available to help me navigate administrative processes.

This dissertation is only a small part of what I have worked on while during my PhD. I am lucky to have taught many exceptional students, to have been part of Duke's vibrant intellectual community, and to have worked with great data science teams at Intact Financial Corporation, American Institutes for Research, Los Alamos National Laboratory, Deepchecks, and Trubrics. I was generously funded by my academic advisors (through the US Department of Energy, Alfred P. Sloan Foundation, National Science Foundation), the Duke University Statistical Science Department, the Natural Sciences and Engineering Research Council of Canada, the Fonds de recherche du Québec - Nature et technologie, and a G-Research award, enabling me

to focus my research on topics that I found to be the most important and neglected in my field.

1. Introduction

This dissertation is broadly about “better data and better models.” We want better data to inform decision and feed into analytical systems, and better models to maximize utility for users.

I focus on statistical components of two specific topics contributing to these goals. The first topic is data linkage (also known as record linkage or entity resolution), which is used to enrich datasets by combining data sources, discovering relationships in data, and ensuring data quality. The second topic is model evaluation (or machine learning system evaluation), which is used to guide model development and selection to maximize value in real-world applications, accounting for multifaceted requirements and uncertainty in performance estimation.

1.1 Motivation

My motivation to work on these two topics comes from a certain level of neglectedness that, combined with their importance and statistical nature, makes them impactful areas of statistical research.

Regarding model evaluation, the increased usage of statistical and machine learning models in a large range of software systems brings a need for an increased focus on quality, testing, and evaluation. In many industries, a new factory floor is the one walked by data scientists and analysts, and on the production line is a stream of models and predictions. This is a new type of manufacturing industry that benefits

from the use of statistical methods for quality control, and that requires specialized statistical tools to measure, test, and evaluate its statistical products.

Regarding better data, the democratized availability of high-performance, general-purpose machine learning models warrants a renewed emphasis on the curation of rich, high-quality, and semantically enriched datasets. In particular, data linkage and entity resolution are used to recover identity relationships that can transform a tabular dataset into much richer knowledge graphs. Links between data points make large datasets more easily navigable for users and provide additional context and features for machine learning models.

1.2 Structure and Overview

Each chapter of this dissertation is a self-contained research article resulting from collaboration between myself and other scientists and stakeholders. The articles are included in chronological order, as this provides insight into the evolution of certain ideas and perspectives. Each includes background information and literature reviews relevant to the topic at hand. In a few sections of the works, collaborators were the main intellectual contributors — these are indicated by the symbol “♠” directly following the section number.

Chapter 2 reproduces Binette and Steorts (2022b) and is about the evaluation of statistical models used for quantifying the prevalence of modern slavery. We propose three methods to assess the reliability of estimates: constructing “ground truth” datasets using an internal consistency approach, analyzing asymptotic convergence and bias due to misspecified assumptions, and using resampling techniques to assess robustness.

Chapters 3 and 4 reproduce Binette et al. (2023) and Binette et al. (2024). They are about methods and software for evaluating entity resolution systems (large-

scale machine learning-based clustering systems). We introduce novel approaches to data labeling, generalization performance estimation, performance monitoring, and error analysis, while accounting for the practical challenges that have hindered the evaluation of entity resolution systems in the past. This is applied to the disambiguation of inventors in U.S. patents data using the public data platform PatentsView.org.

Chapter 5 reproduces Bai et al. (2023), a close collaboration with Eric Bai, that introduces an algorithm for optimizing an F-score objective in bipartite record linkage. We demonstrate how the F-score is a natural objective that can lead to better population size estimates, we provide an efficient implementation of the optimization algorithm, and we demonstrate its use in application to linking U.S. Census data to Union Army records.

Finally, chapter 6 is about broad issues in the evaluation of machine learning systems, and statistical ideas that can be useful in machine learning evaluation practice. We propose the use of an estimands framework to structure model evaluations and their review, showing how it can lead to improved validity and practical usefulness in applications where commonly-used evaluation methodologies fail to properly rank models according to their generalization performance.

2. On the Reliability of Multiple Systems Estimation for the Quantification of Modern Slavery

The quantification of modern slavery has received increased attention recently as organizations have come together to produce global estimates, where multiple systems estimation (MSE) is often used to this end. Echoing a long-standing controversy, disagreements have re-surfaced regarding the underlying MSE assumptions, the robustness of MSE methodology, and the accuracy of MSE estimates in this application. Our goal is to help address and move past these controversies. To do so, we review MSE, its assumptions, and commonly used models for modern slavery applications. We introduce all of the publicly available modern slavery datasets in the literature, providing a reproducible analysis and highlighting current issues. Specifically, we utilize an internal consistency approach that constructs subsets of data for which ground truth is available, allowing us to evaluate the accuracy of MSE estimators. Next, we propose a characterization of the large sample bias of estimators as a function of misspecified assumptions. Then, we propose an alternative to traditional (e.g., bootstrap-based) assessments of reliability, which allows us to visualize trajectories of MSE estimates to illustrate the robustness of estimates. Finally, our complementary analyses are used to provide guidance regarding the application and reliability of MSE methodology.

2.1 Introduction

Modern slavery refers “to situations of exploitation that a person cannot refuse or leave because of threats, violence, coercion, deception, and/or abuse of power” (International Labour Organization, 2017a,b). This term encompasses issues of forced labor, forced sexual exploitation, and forced marriage (International Labour Organization, 2017a). Individuals involved in the recruitment, harboring, and receipt/transportation of victims of such exploitation are referred to as human traffickers (United Nations, 2001, 2000; Sigmon, 2008).¹

The quantification of modern slavery has received increased attention as organizations have come together to produce global estimates (Walk Free Foundation, 2013; Datta and Bales, 2013; International Labour Organization, 2017a; Landman, 2020). These efforts have assisted anti-slavery campaigns, led to increased public awareness, and supported newly implemented governmental policies. One major, recent goal is determining the number of victims of modern slavery. Specifically, in the United Kingdom (UK), the Home Office estimated between 10,000 and 13,000 potential victims of modern slavery in 2013 using multiple systems estimation (MSE) (Silverman, 2014; Bales et al., 2015). Producing this estimate was part of the strategy leading to the UK Modern Slavery Act 2015 (UK Parliament, 2015). Similar studies have been carried out in the Netherlands (van Dijk et al., 2017), in New Orleans (Bales et al., 2019), in the Western United States (U.S.) (Farrel et al., 2019), in Australia (Lyneham et al., 2019), as well as in Serbia, Ireland and Romania (UNODC, 2018a,b,c). These are reviewed in section 2.3.1.

¹ The terms “modern slavery,” “contemporary forms of slavery,” and “human trafficking” are commonly considered synonymous. However, the terminology has been subject to variation and debate due to its history and due to various legal definitions (Feingold, 2010; Chuang, 2014; Cockayne, 2015; Dottridge, 2017; Mende, 2019; Scarpa, 2020; Allain, 2017; Davidson, 2015; Piper et al., 2015; Bunke, 2016). In this paper, we use the term modern slavery as defined by the International Labour Organization and as commonly referred to throughout the literature (section 2.3.1).

MSE, reviewed in section 2.2, is often the only available technique to estimate the prevalence of modern slavery. This is because victims of modern slavery can be out of the reach of traditional surveys. Instead of relying on representative samples, MSE uses case reports from multiple organizations, such as the police and non-governmental organizations, in order to estimate the population size. This approach promises to expose the scale of an issue which could otherwise be ignored.

The foundations of MSE dates back to at least the 17th century, when John Graunt used similar ideas to estimate London’s population in 1661 (Hald, 2005). Laplace later formalized the technique to provide error bounds (Laplace, 1820), Quetelet advocated this approach for more efficient population census (Quetelet, 1827; Stigler, 1986), and these ideas were further developed in Sekar and Deming (1949). Today, MSE (also referred to as capture-recapture) is widespread in population ecology, in epidemiology, for official statistics and in the social sciences (Bird and King, 2018; Bohning et al., 2017).

2.1.1 Two Centuries of Controversy

Despite its potential, MSE has faced harsh criticism for nearly two centuries. Quetelet abandoned MSE following concerns raised by de Keeverberg on the soundness of underlying assumptions (Quetelet, 1827; Stigler, 1986). More recently, Cormack expressed deep concerns regarding the use of similar methods in epidemiological applications (Cormack, 1999a). He stated that “many of these studies give estimates which are not scientifically justified by the underlying data.” This led to an energetic correspondence between Cormack and proponents of MSE (Hook and Regal, 1999; Cormack, 1999b; Hook et al., 2000; Cormack, 2000). In the context of MSE for the quantification of modern slavery, the issues raised by Cormack and the following disagreement was almost exactly repeated in Whitehead et al. (2019) and in the correspondence with ensued with the authors of key modern slavery studies (Vincent et al.,

2020a; Whitehead et al., 2020; Vincent et al., 2020b). The disagreements involved the soundness of assumptions underlying MSE, the robustness of the procedures, and the accuracy of estimates in applications.

2.1.2 Our Contribution

Our goal is to help address these controversies. This is challenging because, as stated by Silverman (2020), “no ‘ground truth’ is available to investigate the accuracy of any estimates.” Instead, we address this issue using the following key observations:

1. data with ground truth can be obtained from available datasets using the internal consistency approach of Hook and Regal (2000); Hook et al. (2012);
2. the convergence and bias of estimates (due to possibly misspecified assumptions) can be characterized in an asymptotic framework; and
3. the reliability of estimates can be diagnosed using resampling techniques.

Regarding (a), we use all publicly available data from past MSE studies on modern slavery (Silverman, 2014; Bales et al., 2015, 2019; van Dijk et al., 2017; Farrel et al., 2019; Lyneham et al., 2019). The internal consistency approach constructs subsets of this data for which ground truth is available, allowing us to evaluate the accuracy of MSE estimators. This approach was recommended in the discussion of Silverman (2020) by Ridout (2020), but has not, to our knowledge, previously been carried out for modern slavery data. Other types of internal validation approaches have been suggested as a fruitful avenue for future research in discussions by Böhning (2020) in Silverman (2020). Regarding (b), we introduce a novel characterization of the large sample bias of estimators as a function of misspecified assumptions. Using our results, we quantify the effect of individual heterogeneity on the bias of estimates. This shows the scale and direction of the bias that can be expected

in reasonable practical situations. Regarding (c), we propose an alternative to traditional (e.g., bootstrap-based) assessments of reliability. Our proposal is more easily interpretable and it involves fewer degrees of freedom in its specification. In practice, it allows us to put modern slavery MSE estimates into the perspective of a hypothetical trajectory of estimates. These trajectories are visualized to showcase the robustness of estimates to small changes in the data. In addition to these three contributions, we provide a thorough review of MSE methodology and a comparison of real data estimates. Our analyses allow us to illustrate obstacles within MSE, to provide practical recommendations, and to provide further directions for research, complementing previous work in this area (Silverman, 2020; Far et al., 2021).

2.1.3 Organization of the Paper

The rest of the paper is organized as follows. Section 2.2 reviews the MSE literature. Section 2.3 describes data from past modern slavery studies, providing comparisons, performing a sensitivity analysis, and discussing MCMC convergence issues with one of the Bayesian models. Section 2.4 evaluates the performance of estimators when ground truth is available using the internal consistency approach. Section 2.5 provides results regarding the bias of estimates, including the consequences of individual heterogeneity. Section 2.6 proposes a visual diagnostic of estimator robustness and showcases its use on modern slavery data. Finally, we discuss main takeaways in section 2.7.

2.2 Background

This section reviews the general framework of MSE. We describe the fundamental idea as reflected in the Lincoln-Peterson estimator in section 2.2.1 and we introduce the general model with multiple lists in section 2.2.2. Section 2.2.3 discusses the assumptions of MSE and their relationship to the existence of consistent population

size estimators. We then review log-linear, decomposable graphical, and latent class models in section 2.2.4.

2.2.1 Introduction to Multiple Systems Estimation

MSE is a technique used to estimate the total size of a population. It relies on multiple samples from the population, referred to as lists, which have been collected by different organizations. The basic principles of the approach are most intuitively explained in the context of two lists. Here two organizations (e.g. the police and a non-governmental organization) record their contact with individuals from the population of interest (such as potential victims of modern slavery). A large overlap between the two lists may indicate that a large portion of the total population was observed. A small overlap may indicate that a larger portion was unobserved. This is justified if the two lists are independent, meaning that the probability that individual appears on one of the lists does not depend on whether an individual appears on the other list.

To formalize this idea, let n_1 be the number of potential victims appearing on a first list, let n_2 be the number on the second list, and let m be the number of potential victims appearing on both. The classical Lincoln-Petersen estimator (Lincoln, 1930; Petersen, 1895) of the total population size is defined as

$$\hat{N} = \frac{n_1 n_2}{m}. \tag{2.1}$$

The estimator is nearly unbiased if the two lists are independent (see Bishop et al. (2007); Chao et al. (2008) for a full account of the properties of the Lincoln-Petersen estimator, extensions, and applications). This was first used to estimate the number of fish in closed reservoirs and extensions led to MSE methodology that is widely used in population ecology (Cormack, 1968; Seber, 1982, 1986, 1992; Amstrup et al., 2005). These extensions have also been adopted in epidemiology (Wittes, 1974; Yip et al.,

1995; Yip et al., 1995; Chao et al., 2001) and to inform public policy (Bird and King, 2018), as well as in human rights applications (Lum et al., 2013; Manrique-Vallier et al., 2013).

2.2.2 The General Framework of Multiple Systems Estimation

Generally, more than two lists are used in MSE studies. This can provide greater coverage of the population of interest and allows the estimation of certain interactions between lists. In this section, we review the general MSE model used for these purposes.

As previously stated, MSE is used to estimate the unknown size N of a population when a complete enumeration is not possible. Instead, $L \geq 2$ lists of observed cases are used to draw inference. Each list records a small fraction of the population. The total number of observed individuals, n_{obs} , is obtained by combining lists and removing any duplicate individuals. The number of unobserved individuals, $n_{\mathbf{0}}$, is estimated using MSE from the patterns of overlap between the lists. Together, the number of observed individuals (n_{obs}) and the number of unobserved individuals ($n_{\mathbf{0}}$) account for the entire population, $N = n_{\text{obs}} + n_{\mathbf{0}}$.

For each individual $i = 1, 2, 3, \dots, N$ in the entire population, we observe a list inclusion pattern $W_i = (W_{i,1}, W_{i,2}, \dots, W_{i,L}) \in \{0, 1\}^L$ where $W_{i,j} = 1$ if individual i appears on list j and $W_{i,j} = 0$ otherwise. If $W_i = (0, 0, \dots, 0) = \mathbf{0}$, then the i th individual is unobserved. The observed data are the counts

$$n_x = \sum_{i=1}^N \mathbb{I}(W_i = x), \quad x \in \{0, 1\}^L \setminus \{\mathbf{0}\}. \quad (2.2)$$

This represents the counts of individuals with given non-zero inclusion patterns. The set of all observed counts is denoted by $(n_x)_{x \neq \mathbf{0}}$ and the number of observed individuals is $n_{\text{obs}} = \sum_{x \neq \mathbf{0}} n_x$.

For example, in the context of the United Kingdom (UK) study (Silverman, 2014; Bales et al., 2015), the lists correspond to organizations coming into contact with potential victims of modern slavery. The organizations are local authorities (LA), the police force (PF), the national crime agency (NCA), governmental organizations (GO), non-governmental organizations (NG), and the general public (GP). Each organization records identifying information for each case it comes into contact with. The resulting lists are then matched together using record linkage or de-duplication (Christen, 2012; Christophides et al., 2021; Binette and Steorts, 2022a), which allows one to identify duplicate records from multiple lists. The observed data, containing all observed overlap counts, is reproduced in table 2.1, where the PF and NCA lists have been combined following Silverman (2014); Bales et al. (2015). In table 2.1, columns under “Cases observed once” represent the number of victims only observed in the list marked by an “×” underneath. The other columns represent the amount of overlap between the lists marked by “×” underneath. For example, 54 potential victims have only been reported by the LA list, that 463 potential victims have only been reported by the NG list, and that 15 potential victims have been reported by both LA and NG but not by any of the other lists. One victim has been observed on LA, NG, PFNCA and GO, but not on GP (rightmost column). In total, 2744 distinct potential victims have been identified in this dataset.

Table 2.1: Counts of potential victims of modern slavery in the UK, disaggregated by the lists in which potential victims appear. This data was reported in Silverman (2014).

	Total	Cases observed once					Cases observed twice								3+ times				
	2744	54	463	995	695	316	15	19	3	62	19	1	76	11	8	1	1	4	1
LA	×						×	×	×							×	×		×
NG		×					×			×	×	×				×	×	×	×
PFNCA				×				×		×			×	×		×		×	×
GO					×				×			×			×		×	×	×
GP						×						×		×	×				

2.2.3 Assumptions and Consistency of Estimators

This section reviews the three assumptions of the standard MSE model regarding list inclusion patterns. The first two are standard, stating the data is independently and identically distributed. The third is an identifiability assumption which we show in Proposition 1 is necessary to the existence of consistent population size estimators.

The first two assumptions are formally stated below:

A1 The list inclusion patterns W_i , $i = 1, 2, 3, \dots, N$ are independent from one another. That is, the lists on which individuals appear or do not appear has no influence on the inclusion patterns of other individuals — this is independence across individuals.

A2 The list inclusion patterns W_i , $i = 1, 2, 3, \dots, N$ are identically distributed with

$$p_x = \mathbb{P}(W_i = x) > 0, \quad x \in \{0, 1\}^L. \quad (2.3)$$

That is, for any set of lists represented by an inclusion pattern $x \in \{0, 1\}^L$, all individuals are equally likely to have x as an inclusion pattern.

A1 and **A2** are classical assumptions in the MSE literature and we point the reader to (Lum et al., 2013) for a practical discussion of MSE assumptions. While assumption **A1** may not always hold, we expect that it has a non-negligible effect for large populations. Assumption **A2** is also less stringent than it appears. It only requires the data to be marginally identically distributed. That is, suppose that the inclusion pattern probability of an individual i is affected by an unobserved variable λ_i , which represents the type of crime involved or socio-demographic characteristics of the individual victim. As long as $\mathbb{E}_{\lambda_i} [\mathbb{P}(W_i = x \mid \lambda_i)] = p_x$ is constant and does not depend on i , then assumption **A2** is satisfied. Such models, where inclusion probabilities depend on latent individual characteristics, are referred to as “individual

heterogeneity” models (Otis et al., 1978). These are relevant to modern slavery applications given the heterogeneity in the population which may impact list inclusion probabilities. Throughout the paper, we assume that **A1** and **A2** hold. Heterogeneity models are specifically considered in section 2.5.2.

Here, we introduce a decomposition of the data likelihood due to Fienberg (1972). Under assumptions **A1** and **A2**, the observed data $(n_x)_{x \neq \mathbf{0}}$ is distributed as

$$n_{\text{obs}} \sim \text{binomial}(1 - p_{\mathbf{0}}, N), \quad (2.4)$$

$$(n_x)_{x \neq \mathbf{0}} \mid n_{\text{obs}} \sim \text{multinomial}((q_x)_{x \neq \mathbf{0}}; n_{\text{obs}}), \quad (2.5)$$

where

$$q_x = p_x / (1 - p_{\mathbf{0}}) = \mathbb{P}(W_i = x \mid W_i \neq \mathbf{0}), \quad x \neq \mathbf{0}$$

is the conditional probability of the inclusion pattern x given the individual being observed. In some cases, a Poisson likelihood is used as an approximation to the multinomial model (2.4) and (2.5) (Cormack, 1989). This is the case with the Poisson log-linear modeling approach reviewed in section 2.3.2. The Poisson likelihood is used for convenience and does not make any important difference on the model, its assumptions, or resulting estimates.

While the population size N is identifiable in the standard MSE model given by (2.4) and (2.5) (Farcomeni and Tardella, 2012), assumptions **A1** and **A2** are not sufficient by themselves to obtain meaningful population size estimates. Indeed, it follows from (2.4) that the observed data provides information about N only through n_{obs} and $p_{\mathbf{0}}$. In section 2.5, we formally define the notion of consistent population size estimators to formalize what can be learned from the data. Roughly, a population size estimator \hat{N} is consistent for a statistical model $\Theta \subset \{(p_x)_{x \neq \mathbf{0}} : \sum_x p_x = 1\}$ if and only if it converges to the true population size N in large samples whenever the probabilities p_x in (2.3) are part of Θ . The minimal requirement for the existence of consistent population size estimators is given in Proposition 1 below.

Proposition 1. *If a model Θ admits a consistent population size estimator, then there exists a function f such that $p_0 = f((q_x)_{x \neq 0})$ for all $(p_x)_{x \in \{0,1\}^L} \in \Theta$, where $q_x = \frac{p_x}{1-p_0}$.*

The proof is in Appendix 2.8.3.

Proposition 1 motivates assumption **A3** which, given some function f , restricts the set of probabilities (p_x) under consideration.

A3 There exists a function f such that

$$p_0 = f((q_x)_{x \neq 0}), \quad \text{where} \quad q_x = \frac{p_x}{1-p_0}.$$

That is, the unobserved probability p_0 is the deterministic function f of the observed data distribution through the cell probabilities $(q_x)_{x \neq 0}$.

We refer to this as the identifying assumption, which formalizes a condition of Link (2003). That is, **A3** is equivalent to stating that no two different distributions $(p_x)_{x \in \{0,1\}^L}$ lead to the same zero-truncated distribution $(q_x)_{x \neq 0}$. Our assumption **A3** is equivalent to Definition 1 in Aleshin-Guendel (2020) and Definition 2 in Aleshin-Guendel et al. (2021). It relates part (2.5) of the observed data distribution to the probability of an individual being unobserved. As such, it can be interpreted as specifying the missing data mechanism at play in MSE. It can also be understood as an extrapolation formula (Manrique-Vallier et al., 2021), which allows one to go from the observed data distribution $(q_x)_{x \neq 0}$ to the unobserved probability p_0 . Then through (2.4), one can infer the total population size.

An example of an assumption of the form **A3** is given in section 2.2.4.1 in the context of log-linear modeling.

Remark 1. The observed data provides no information regarding assumption **A3**. In fact, as pointed out by Manrique-Vallier et al. (2021), “the way in which the

probability $p_{\mathbf{0}}$ relates to the rest of p_x , $x \neq \mathbf{0}$, can neither be learned from data nor tested.” Turning the choice of f in **A3**, in practice, this is taken as the consequence of simpler and more easily interpretable assumptions. Typically, f is chosen as the consequence of one of the particular modeling approaches described next section 2.2.4.

2.2.4 Multiple Systems Estimation Methods

In this section, we review MSE models widely used in the literature. These allow one to estimate the probabilities p_x defined in (2.3) and to specify a function f in **A3**, that provides via (2.4) a population size estimate. Crucially, we only consider models that do not inherently require covariate-level data as motivated by the applications in section 2.3.1. First, we review log-linear models (Fienberg, 1972; Cormack, 1989). Second, we review graphical models (Madigan et al., 1995; Madigan and York, 1997), which are a special case of log-linear models. Third, we review a family of latent class models (Manrique-Vallier and Fienberg, 2008; Manrique-Vallier, 2016; Aleshin-Guendel, 2020).

2.2.4.1 Log-Linear Models

Fienberg (1972) introduced the use of log-linear models for MSE, which provide the basis for many applications (Yip et al., 1995; Baillargeon et al., 2007). Furthermore, the use of these models have been previously considered in the use of modern slavery studies (section 2.3) (Silverman, 2014; Bales et al., 2015, 2019; Silverman, 2020; Chan et al., 2020; van Dijk et al., 2017; Farrel et al., 2019; Lyneham et al., 2019). In this section, we briefly review this literature, referring the reader to (Bishop et al., 2007; Yip et al., 1995) for further details.

Log-linear models provide an interpretable re-parameterization of the model parameters p_x and N defined in (2.3). The log-linear model consists of an intercept term μ , a main list effects term α_i , two-way interaction terms $\beta_{i,j}$ (for lists $i \neq j$),

and higher order interaction terms including a full-way interaction term γ . Thus, the log-linear parameterization of N and $(p_x)_{x \in \{0,1\}^L}$ is given by

$$\log(Np_x) = \mu + \sum_i x_i \alpha_i + \sum_{i \neq j} x_i x_j \beta_{i,j} + \cdots + x_1 x_2 \cdots x_L \gamma, \quad x \in \{0,1\}^L. \quad (2.6)$$

The full-way interaction term γ can be expressed as

$$\gamma = \sum_{x \in \{0,1\}^L} (-1)^{|x|+1} \log(p_x) \quad (2.7)$$

where $|x| = \sum_i x_i$.

We now discuss the parameter interpretation in (2.6). If all parameters other than μ and the α_i are zero, this corresponds to independent lists. The probability that an individual appears in list i is given by $e^{\alpha_i}/(1 + e^{\alpha_i})$. If two-way interaction terms are added, then $\beta_{i,j}$ represents how the odds of inclusion to non-inclusion on the i th list, conditionally on all other variables, changes depending on whether or not the individual appears on the j th list. For non-overlapping lists, setting $\beta_{i,j} = -\infty$ states that an individual cannot appear on both lists i and j . Higher order interaction terms can be similarly interpreted in terms of log-odds changes.

One of the main advantages of the log-linear parameterization is the resulting model hierarchy. The independence model, with intercept μ and main effects α_i , is obtained by setting all interaction terms to zero. More complex models can be obtained by adding interaction terms, allowing for dependencies between lists. Typically, simpler models are favored, with interaction terms only added to the extent that data provides evidence for them. This has been called “betting on sparsity” (Friedman et al., 2001). For instance, all log-linear models can be fitted to the data using maximum likelihood estimation, and a single model can be selected based on Akaike’s Information Criteria or through other criteria (Chao et al., 2001). Regal and Hook (1991); Yip et al. (1995) review other considerations involved in the selection of log-linear models and the reporting of corresponding estimates.

Crucially, the assumption of no full-way interaction in log-linear models, which corresponds to setting $\gamma = 0$ in (2.6), induces an identifying assumption of the form **A3**. The following assumption is typically made within the context of log-linear models (Fienberg, 1972):

A3.1 The full-way interaction term γ in the log-linear parameterization (2.6) is zero. Equivalently, the probabilities (p_x) defined in (2.3), with $q_x = p_x/(1 - p_{\mathbf{0}})$, satisfy the relationship

$$\log \frac{p_{\mathbf{0}}}{1 - p_{\mathbf{0}}} = \sum_{x \neq \mathbf{0}} (-1)^{|x|} \log q_x. \quad (2.8)$$

This states that the $(L - 1)$ -way interaction term for individuals appearing on list L is the same as the $(L - 1)$ -way interaction term for individuals not appearing on list L .

Remark 2. As discussed in section 2.2.3, assumption **A3.1** is untestable and is made in order to obtain consistent population size estimators. Yip et al. (1995) states that this assumption “is more likely to be approximately correct than assumptions about the absence of lower-order interactions.” In the case of only two lists, **A3.1** is satisfied if the two list inclusion indicators W_1 and W_2 are uncorrelated or independent. With more than two lists, if one list is independent of another given the rest, then **A3.1** is also satisfied.

2.2.4.2 Decomposable Graphical Models

We now review decomposable graphical models (Darroch et al., 1980) with hyper-Dirichlet priors (Dawid and Lauritzen, 1993), which were proposed by York and Madigan (1992); Madigan et al. (1995); Madigan and York (1997) for MSE. First, we describe (undirected) graphical models, which are special cases of log-linear models with an intuitive interpretation of conditional dependencies. Second, we review

decomposable graphical models with hyper-Dirichlet priors, which are mainly used for computational convenience as they provide a conjugate family for which population size estimates can be derived in closed form.

2.2.4.2.1 Graphical Models

Graphical models are statistical models where the dependency between variables is characterized by an interpretable graph. Nodes in the graph represent variables, and two variables are linked together if they exhibit certain conditional dependencies. In the context of MSE, the graph describes the dependencies and conditional independencies between lists. There is one node for each list; it represents the variable indicating whether or not a given individual appears on this list. The graph has two equivalent properties, known as the local Markov property or the conditional independence graph, which we define below.

Local Markov property: The conditional distribution of any variable only depends on other variables through its immediate neighbors in the graph.

Conditional independence property: If A and B are two sets of vertices separated by another set S in the graph, then the variables corresponding to A and B are conditionally independent given S .

Figure 2.1 presents examples of conditional independence graphs for the list considered in the UK modern slavery study of Silverman (2014); Bales et al. (2015).

2.2.4.2.2 Decomposable Graphical Models and the Hyper-Dirichlet Prior

In this section, we review other important terminology used in graphical models. Decomposable graphical models refer to graphical models for which the graph G is chordal — every cycle in the graph is part of a clique.² From a statistical point of view,

² A cycle is a set of vertices which are connected in a closed chain. A clique is a set of vertices which are all interconnected.

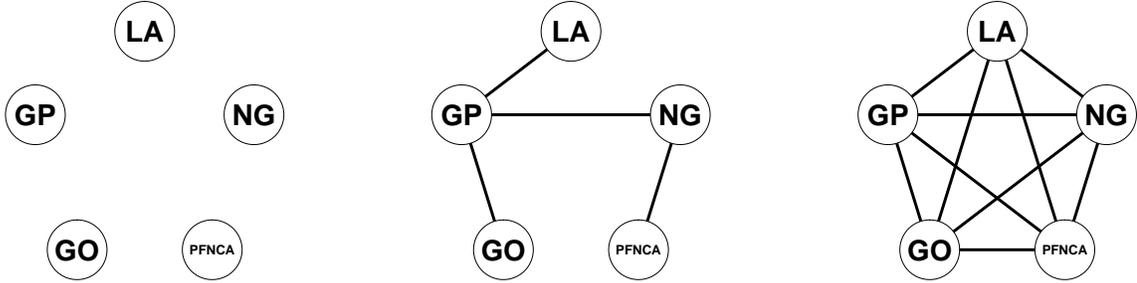


FIGURE 2.1: Examples of independence graphs on the 5 lists considered in the UK study of Silverman (2014); Bales et al. (2015). **Left:** Full independence between the lists. **Middle:** Various dependencies and conditional independencies. For example, NG is conditionally independent of GO and LA given GP. **Right:** Full dependency model with no non-trivial conditional independencies.

decomposable graphical models have an important advantage in that every distribution on a decomposable graph is uniquely characterized by the set of marginal distributions over its cliques. That is, no two different distributions on a decomposable graph can have the same marginal distributions over the cliques of the graph. Furthermore, any set of pairwise consistent marginal distributions over the cliques of a decomposable graph uniquely determines a joint distribution. Here, distributions on two sets of vertices A and B are said to be consistent if they have the same marginal on $A \cap B$. In other words, probability distributions over decomposable graphs can be specified through clique marginals which are pairwise consistent.

Dawid and Lauritzen (1993) exploited these properties to define the hyper-Dirichlet prior, a prior distribution for decomposable graphical models. This prior distribution is easily specified through Dirichlet clique marginals, allowing for tractable posterior inference.

2.2.4.2.3 Bayesian Model Averaging of Decomposable Graphical Models

In the context of MSE and of Madigan and York (1997), the hyper-Dirichlet prior on decomposable graphical models is parameterized by a set of “prior counts” α_x for $x \in \{0, 1\}^L$ (Sadinle, 2018). For a given decomposable graph, the marginalization of

these counts over the cliques of the graph yields the corresponding parameters of the marginal Dirichlet priors. Given a prior on the population size, we obtain a tractable formula for the posterior distribution (see Madigan and York (1997); Sadinle (2018)). Uncertainty regarding the structure of the decomposable graph is incorporated using a prior on the set of all possible graphs, known as Bayesian model averaging (Gelman et al., 2013).

Excluding the complete graph, all graphical models are special cases of log-linear models with no full-way interactions (Darroch et al., 1980). As such, graphical modeling relies on assumption **A3.1** of no full-way interaction.

2.2.4.3 Latent Class Models

Finally, we review the latent class model of Manrique-Vallier (2016), which is motivated by the modeling of latent individual heterogeneity.

The latent class model decomposes the probabilities p_x defined in (2.3) through a mixture representation. That is, the model takes the form

$$p_x = \sum_{k=1}^K w_k \prod_{j=1}^L \lambda_{k,j}^{x_j} (1 - \lambda_{k,j})^{1-x_j} \quad (2.9)$$

where w_k is the class weight and where $\lambda_{k,j}$ is the j th list inclusion probability for class k . Any distribution $(p_x)_{x \in \{0,1\}^L}$ can be decomposed through (3) with $K = 2^{L-1}$ (Johndrow et al., 2017). In terms of the list inclusion variables, this model represents independence between lists conditionally on the latent class to which belongs an individual.

Aleshin-Guendel (2020); Aleshin-Guendel et al. (2021) showed model (2.9) induces an assumption of the form **A3** if and only if $2K \leq L$. When $2K > L$, it follows from Aleshin-Guendel (2020) and Proposition 1 that no consistent population size estimator results from (2.9) with unrestricted values of w_k and $\lambda_{k,j}$. In other words, given a

prior on the latent class model, the resulting population size posterior distribution is generally inconsistent — the posterior distribution does not converge to the true population size. Instead, such an approach quantifies uncertainty through the prior relationship between p_0 and the other model probabilities. This is valid as long as the mixture of independence model is appropriate for the data, and as long as the prior specification properly captures our uncertainty regarding the distribution of the latent classes and the list inclusion probabilities. The prior specification of Manrique-Vallier (2016) for (2.9) is discussed in Section 2.3.2.

2.3 Data and Population Size Estimators Under Consideration

This section introduces the datasets and estimators which we consider throughout the paper. Section 2.3.1 reviews the datasets and section 2.3.2 introduces the estimators based on the models of section 2.2.4. In section 2.3.3, we then showcase how the corresponding estimates compare among themselves and in relation to published estimates. Finally in section 2.3.4, we discuss challenges associated with these approaches, notably the sensitivities to choices of hyperparameters and convergence issues.

2.3.1 Data From Past Modern Slavery Studies

In this section, we summarize all publicly available datasets from past studies on modern slavery that utilized MSE (table 2.2). The United Kingdom, New Orleans, Netherlands, and Western U.S. datasets were considered in a recent study of Silverman (2020). All data considered has been stripped of covariate information, although in some of the original studies (Netherlands and Western U.S.) such covariate data was available to the researchers and used to produce estimates. Furthermore, we only consider a maximum of five lists for each of the datasets, in order to ensure

Table 2.2: Datasets under consideration, the timeframe for the collected data, the number of observations, the number and proportion of overlap (observations which appeared in more than one list), and the total number of lists. In order from top to bottom, datasets are from Silverman (2014), Bales et al. (2019), van Dijk et al. (2017), Farrel et al. (2019), and Lyneham et al. (2019).

Dataset	Timeframe	# observations	# overlap	# lists
United Kingdom	2013	2744	221 (8.1%)	5
New Orleans	2016	185	12 (6.5%)	5
Netherlands	2010–2015	8234	431 (5.2%)	5
Western U.S.	2016	345	23 (6.7%)	5
Australia	2015–16 to 2016–17	414	69 (16.7%)	4

applicability of the decomposable graphical model approach of Madigan and York (1997); Lum et al. (2015). For the United Kingdom, New Orleans, and Netherlands datasets which originally contained more than five lists, we consider the five lists version proposed in Silverman (2020). Full details on each dataset can be found in Appendix 2.8.1.

2.3.2 Population Size Estimators

We now introduce the population size estimators which we evaluate and compare. The choice is motivated by past MSE studies for the quantification of modern slavery, as well as by the recent comparative analysis of Silverman (2020). First, we consider the approach of Chan et al. (2020), which we refer to as SparseMSE following the name of the corresponding R package. This approach was motivated by the studies of Silverman (2014); Bales et al. (2015, 2019). It addresses the issues of non-overlapping lists and of model selection uncertainty. Second, we consider the approach of Madigan and York (1997), which we refer to as dga following its implementation in the dga R package of Lum et al. (2015). This approach was considered in Silverman (2020) and provides a Bayesian model averaging approach to MSE. Third, we consider the approach of (Manrique-Vallier, 2016) which we refer to as LCMCR following the name

of the corresponding R package. This estimator was also considered in Silverman (2020) and has been used in human rights statistics, leading to recent extensions (Kang et al., 2020). Finally, we consider a simple independence model as a baseline point of reference. Each approach provides point and interval estimates, relying on the modeling approaches reviewed in section 2.2.4 for model fitting and inference.

SparseMSE: SparseMSE (Chan et al., 2020) fits a log-linear model of the form (2.6) with no three-way or higher interaction terms and with two-way interaction terms selected through forward stepwise p -value thresholding. A Poisson likelihood approximation to the multinomial data likelihood is used for mathematical convenience; this does not meaningfully change inferences (Cormack, 1989). Extended maximum likelihood is used for parameter estimation, accounting for non-overlapping lists. The “bias-corrected and accelerated” bootstrap procedure of DiCiccio and Efron (1996) is used for the construction of confidence intervals while accounting for model selection uncertainty. Crucially, SparseMSE relies on the assumption of no full-way interaction, meaning that $\gamma = 0$ in the log-linear parameterization (2.6) of the model probabilities.

dga: The dga approach (Madigan and York, 1997; Lum et al., 2015) uses decomposable graphical models with hyper-Dirichlet priors and Bayesian model averaging, as described in section 2.2.4.2, to obtain a population size posterior distribution. By default, “prior counts” are set to be constant and with value 2^{-L} , where L is the number of lists, the prior on the set of decomposable graphs is constant, and the population size prior is the improper prior $p(N) \propto 1/N$. As a population size estimator, we consider the median of the posterior distribution. Confidence intervals are obtained by taking equally tailed quantiles of the posterior distribution. This approach also makes the assumption of no full-way interaction, meaning that $\gamma = 0$ in the log-linear parameterization (2.6) of the

model probabilities.

LCMCR: LCMCR (Manrique-Vallier, 2016) uses the latent class representation (2.9) together with a stick-breaking prior on the weights w_k and a uniform prior on the list inclusion probabilities $\lambda_{k,j}$. The stick-breaking prior defines $w_1 \sim \text{Beta}(1, \alpha)$, $w_2 = (1 - w_1)v_2$ with $v_2 \sim \text{Beta}(1, \alpha)$, $w_3 = (1 - w_2)v_3$ with $v_3 \sim \text{Beta}(1, \alpha)$, and so forth, with α itself being Gamma distributed. By default, $\alpha \sim \text{Gamma}(0.25, 0.25)$. The population size N is given the default improper prior $p(N) \propto 1/N$. The posterior distribution of N is approximated through conjugate Gibbs sampling. By default, we run 200 randomly initialized chains, each with 100,000 iterations which are thinned down to 100 samples. The number of latent classes is limited to a maximum number of 10 classes to reduce computational burden. We summarize the population size posterior distribution using the posterior median and equally tailed quantiles for confidence intervals.

Independence: Additionally, we consider an independence model as a baseline point of reference. This is a Poisson log-linear model with no two-way or higher interaction terms. It is fitted to the data through maximum likelihood, using the `modelfit()` function of the `SparseMSE` R package (Chan et al., 2020).

Remark 3. There are many other models and estimators used for capture-recapture and multiple systems estimation (Otis et al., 1978; Amstrup et al., 2005; Baillargeon et al., 2007; Laake et al., 2013; Overstall and King, 2014; Bohning et al., 2017; Worthington et al., 2021). Our paper focuses on approaches which have previously been used in multiple systems estimation studies for the quantification of modern slavery and which are suited to modern slavery data. In comparison, many capture-recapture models from the population ecology literature require some amount of experimental control to justify strong underlying assumptions such as assumptions of independence between lists. To our knowledge, experimental control is not present in

the modern slavery applications that we consider, which could make the modeling assumptions ill-suited. Thus, we focus on previously-used techniques with more realistic assumptions.

2.3.3 Comparison of Estimates

Figure 2.2 shows the comparison of the SparseMSE, dga, LCMCR, and Independence estimates on the datasets introduced in Section 2.3.1.

All of the approaches considered, except for the independence model, account for model selection uncertainty, whereas published estimates from past studies did not. This explains the very narrow uncertainty in some recent published estimates when compared to the estimates of the SparseMSE, dga, and LCMCR approaches.

Observe that there is general agreement between the estimates. This is particularly pronounced in the case of the New Orleans and Western U.S. datasets for which little overlap between lists is available in the data. Only 12 cases appeared in more than one list in New Orleans, and only 23 cases appeared on more than one list the Western U.S. Without overlap in the data to accurately estimate interaction terms, the regularization implicit to these approaches tends to produce estimates that are in alignment with the independence model estimates.

Overall, the dga and LCMCR estimates tend to be comparable. This is not something which should be expected given that the dga and LCMCR models rely on different identifying assumptions. However, both models contain the independence model as a particular case. The independence model estimates are similar to most other estimates, although it provides very narrow confidence intervals on larger datasets. The SparseMSE estimates notably differ in the cases of the United Kingdom and of Australia.

In section 2.3.4, we explore some of the sensitivities of these approaches to tuning parameters which may influence the results. The sensitivity analysis highlights

challenges associated with the use of each estimator.

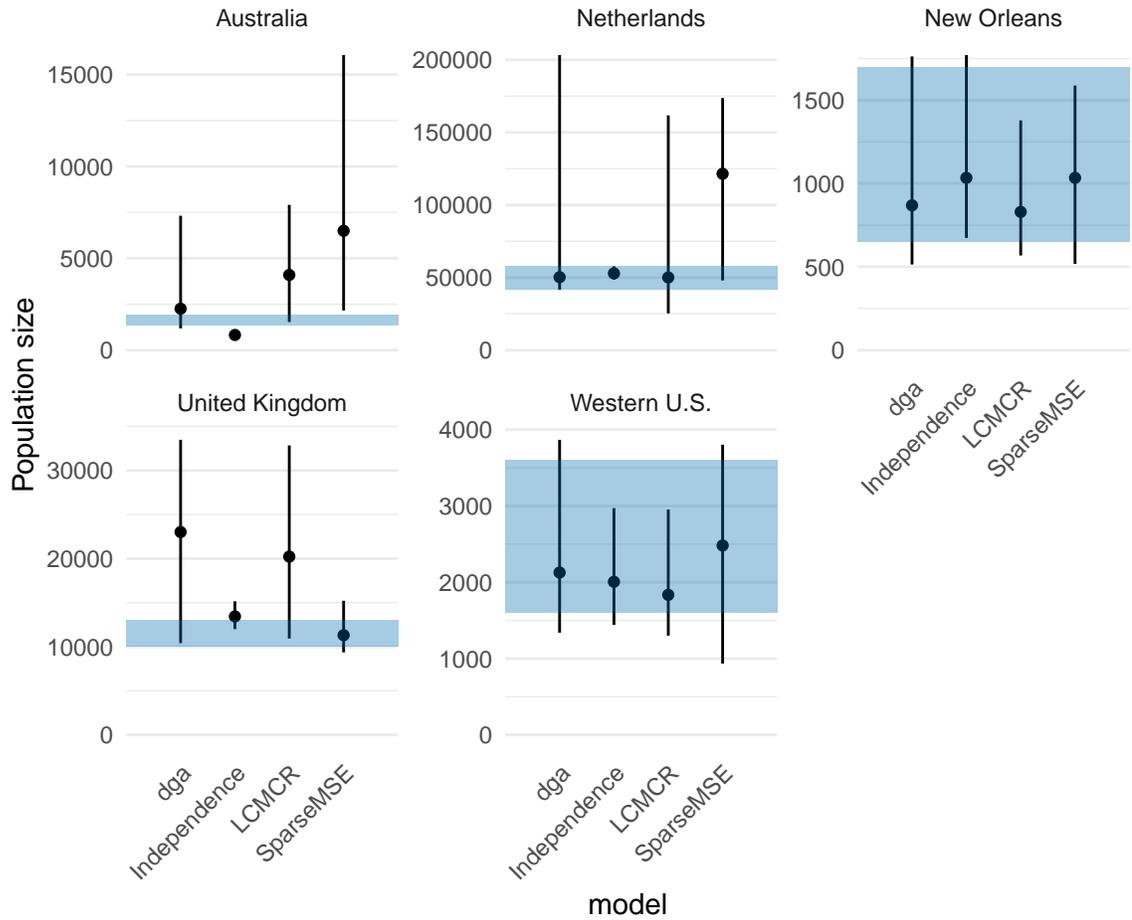


FIGURE 2.2: Comparison of the independence model, SparseMSE, LCMCR and dga estimates on the modern slavery datasets described in Section 2.3.1. Vertical line ranges represent 95% confidence intervals. The blue shaded regions represent the published estimates.

2.3.4 Sensitivity and Convergence Issues

Let us now turn to the sensitivities of the SparseMSE and dga estimates to choices of tuning parameters, as well as convergence issues with LCMCR. The sensitivities and convergence issues can be major challenges in practice, even before potential issues with the data and models are considered. Note that we focus on convergence issue

with LCMCR, rather than its sensitivity to choices of priors, as this is, we believe, the most important problem which the approach currently faces. We also ignore the Independence estimator which requires no tuning parameters.

Throughout, we focus on highlighting issues on datasets for which they are most noticeable or most relevant. Our goal is to showcase issues which can happen and which should be evaluated in practice, rather than to provide an analysis for each of the five considered datasets.

2.3.4.1 Sensitivity of SparseMSE Estimates

We consider the main tuning parameter of SparseMSE — the p -value threshold used for stepwise model selection. Bales et al. (2015) used a threshold of 0.05 and Chan et al. (2020) used a threshold of 0.02. Intuitively, we would expect smaller threshold to lead to bias towards the independence model, whereas a higher threshold allows for the consideration of more complex models with correspondingly higher estimator variance. This is not necessarily the case. Figure 2.3 showcases the SparseMSE estimates on the United Kingdom dataset for p -value thresholds between 0 and 0.1. While the threshold of 0.02, used in Chan et al. (2020), led to the narrow confidence interval of between 10,000 and 15,000 potential victims, a very slightly smaller threshold leads to the much bigger interval of between 10,000 and 30,000 potential victims. This is surprising behavior - smaller thresholds should correspond to higher regularization, but they can unexpectedly produce much larger confidence intervals due to higher model selection uncertainty.

In practice, larger p -value thresholds might be preferable given that they allow the estimation of more complex interactions between lists. However, if there is no strong justification for the use of one p -value threshold over another, we recommend that the range of estimates corresponding to different thresholds be investigated and reported in applications.

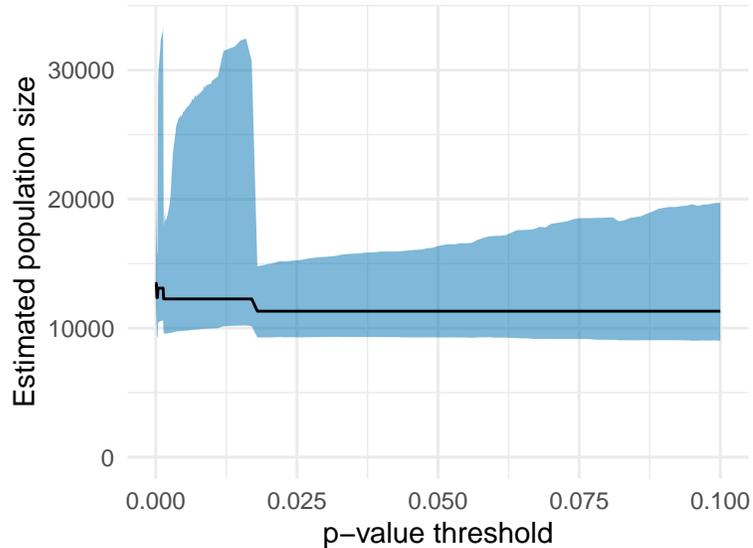


FIGURE 2.3: Estimates of the SparseMSE approach on the United Kingdom data, for p -value thresholds between 0 and 0.1. The black line represents the point estimates and the blue band represents the 95% bootstrap confidence intervals.

2.3.4.2 Sensitivity of dga Estimates

Let us now turn to dga estimates. We explore their sensitivity to reasonable choices of hyperparameters or prior distributions. There are two prior distributions for which it may be difficult to elicit informative priors and which we focus on. Specifically, we consider the choice of “prior counts” which determine the hyper-Dirichlet prior on decomposable graphs and the choice of prior on the set of decomposable graph structures.

First, regarding prior counts, Lum et al. (2015) proposes the default $\delta = 2^{-L}$ where L is the number of lists. The choice of $\delta = 2^{-L}$ corresponds to the expected count under an independence model for which each list has an inclusion probability of $1/2$. We extend this prior by considering the expected count under an independence model where each list has an inclusion probability $\kappa > 0$. Values $\kappa < 0.5$ corresponds to lower probabilities of inclusion on individual lists, which seems more reasonable in

the context of modern slavery data.

Second, regarding the prior on the graphical structure, we consider a “small-world” prior restricted to decomposable graphs. That is, each edge in the graph appears with independent probability $\beta \in (0, 1)$, conditionally on the resulting graph being decomposable (and with the complete graph being excluded as well). Larger values of β gives more weight to more complex models, while smaller give more weight to less complex models. The default uniform prior corresponds to setting $\beta = 1/2$.

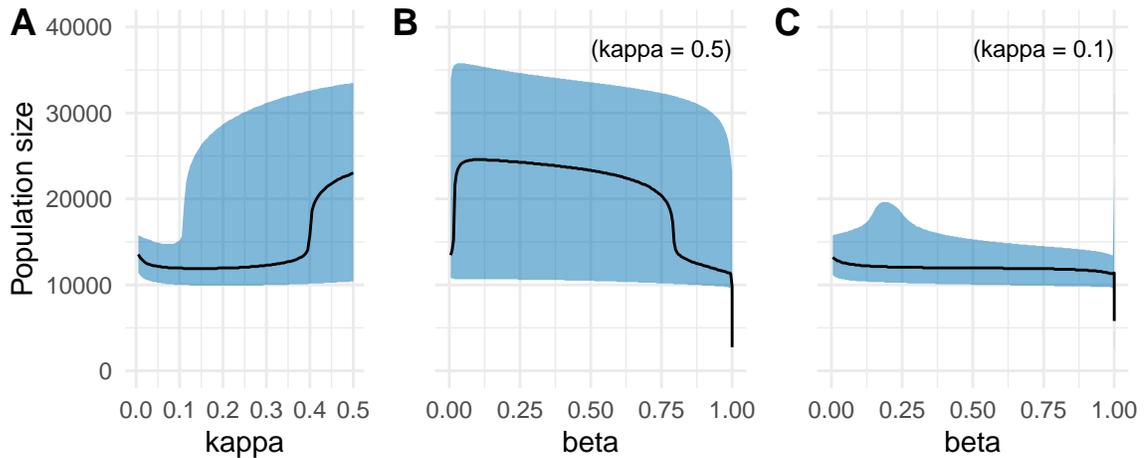


FIGURE 2.4: Point estimates and 95% credible intervals of dga estimates for the United Kingdom dataset, with varying values of κ (which parameterizes the prior counts) and β (which parameterizes the distribution on graphical structures). In panel **A**, β is fixed to a value of $1/2$ and κ ranges between 0 and 0.5. In panel **B**, $\kappa = 0.5$ and β ranges between 0 and 1. In panel **C**, $\kappa = 0.1$ and β ranges between 0 and 1.

Figure 2.4 shows estimates resulting from different choices of κ and β in application to the United Kingdom dataset. We see that prior choices can quite heavily influence estimates and the width of confidence intervals. In practice, unless a particular prior can be rigorously justified, estimates should be reported for the whole range of plausible prior distributions. This can be viewed as part of an “objective Bayesian analysis,” where we acknowledge the difficulty of selecting a prior distribution.

2.3.4.3 LCMCR Convergence Issues

Like any other Bayesian approach, LCMCR estimates are sensitive prior choices and ranges of reasonable priors should be explored in practice. However, LCMCR faces an important additional challenge – the MCMC algorithm used to compute estimates does not converge in some cases. This problem is due to the non-identifiability of the latent class model which results in a multimodal posterior distribution. With large datasets, such as the Netherlands data, the Gibbs sampling algorithm of Manrique-Vallier (2016) can struggle to explore the posterior distribution.

Figure 2.5 provides the Markov chain Monte Carlo (MCMC) samples used to approximate the LCMCR posterior distribution of the non-observation probability in application to the Netherlands dataset. Observe one trace plot for 200 independent chains, where each chain was run for 100,000 iterations and thinned down to 100 samples. We find that there are two posterior modes and a lack of mixing between them. We provide MCMC convergence diagnostics in table 2.3 (left) for the non-observation probability p_0 , for the number of unobserved individuals n_{obs} , and for the number of latent classes k^* . The \hat{R} value (Carpenter et al., 2017; Gelman et al., 2013) of 1.67 for the non-observation probability, as well as the effective sample size n_{eff} of only 340 for the total number of 20,000 samples across chains, is witness to non-convergence. Using 20 chains, each running 1000 times longer and thinned down to 1,000 samples, results in a lower effective sample size. Anecdotally, we have not been able to run the Gibbs sampler long enough to observe proper mixing of the non-observation probability.

Given this lack of convergence, we can use a large number of randomly initialized parallel chains to ensure stability of estimates across replications. This explains our default choice of 200 independent chains in our analyses. Other more sophisticated approaches can be used to deal with peaked and multimodal posteriors which mix

Table 2.3: Convergence diagnostics for LCMCR samples aggregated across chains, both for our default settings (left) and for 20 chains each run 1000 times longer than by default (right). Here n_0 represents the number of unobserved individuals, p_0 is the non-observation probability, and k^* is the number of latent classes. The large \hat{R} values and low effective sample sizes are indicative of poor MCMC mixing.

200 chains of 10^5 iterations				20 chains of 10^8 iterations			
	\hat{R}	n_{eff}	n_{samples}		\hat{R}	n_{eff}	n_{samples}
n_0	1.67	340.84	$2 \cdot 10^4$	n_0	1.46	38.04	$2 \cdot 10^5$
p_0	1.67	340.48	$2 \cdot 10^4$	p_0	1.46	38.03	$2 \cdot 10^5$
k^*	1.39	455.64	$2 \cdot 10^4$	k^*	1.01	2338.91	$2 \cdot 10^5$

poorly, such as parallel tempering (Earl and Deem, 2005), using parallel chains (Gelman et al., 1992), and more (Yao et al., 2020). Implementing these approaches would be necessary for the application of LCMCR to larger datasets. We only encounter this issue for the Netherlands dataset, and given the scope of our paper, we leave this for future work.

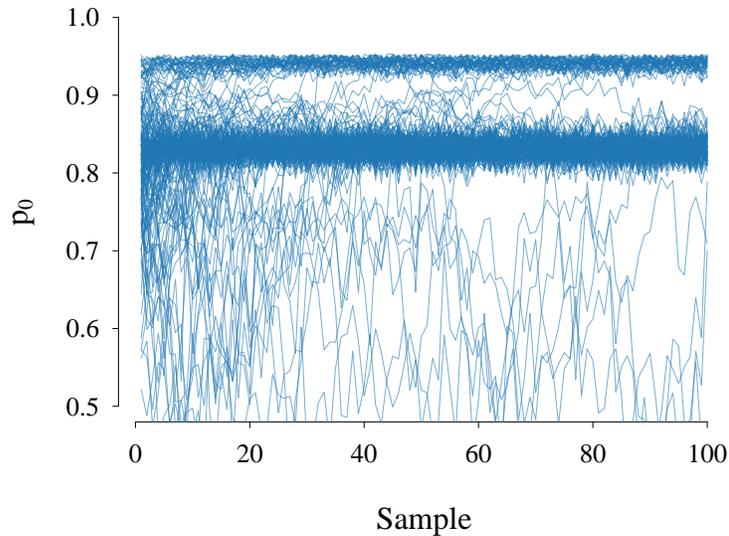


FIGURE 2.5: MCMC traces of the non-observation probability p_0 for 200 independent chains using the Gibbs sampler of Manrique-Vallier (2016), applied to the Netherlands dataset.

2.4 Internal Consistency Analysis

We now turn to our first analysis of the accuracy of MSE estimates in application to data from modern slavery studies. This analysis relies on subsets of the data for which “ground” truth is available, which means that the true population size is already known. This was termed an “internal consistency analysis” by Hook et al. (2012) (see also Hook and Regal (2000); Brittain and Böhning (2009)). This provides a way to evaluate the accuracy of MSE on relevant datasets. The way in which ground truth data is obtained is described in section 2.4.1 and the performance of MSE estimators on this data is described in section 2.4.2. Limitations of this approach which motivate the rest of our paper are discussed in section 2.4.3.

2.4.1 Ground Truth Data Through Conditioning

To illustrate how we obtain data with ground truth, consider the United Kingdom dataset reproduced in table 2.1. This dataset contains five lists, including the local authorities (LA) list. Conditioning on cases being recorded by the LA list and omitting the LA list itself, we obtain the conditioned data shown in table 2.4. In addition, we know that a total of 94 cases have appeared on the LA list. We may therefore attempt to use the conditioned data, which record 40 cases having appeared on the LA list as well as on other lists, in order to estimate the total of 94 cases which appeared on the LA list. Here the LA list is our reference list, and 94 is the ground truth population size for the conditioned data.

2.4.2 Analysis and Results

The process of using a reference list to obtain conditioned data and a corresponding ground truth is repeated for every dataset in table 2.2 and for every list. Datasets with fewer than 30 observations are discarded, resulting in the total of 11 conditioned

Table 2.4: United Kingdom dataset conditioned on the LA list. Note that no cases appeared on both the LA list and the GP lists.

	15	19	3	1	1	1
NG	×			×	×	×
PFNCA		×		×		×
GO			×		×	×
GP						

Table 2.5: Description of conditioned datasets with more than 30 observations.

Dataset	Reference list	Ground truth	# observations	# overlap
United Kingdom	LA	94	40	3
	NG	567	104	7
	PFNCA	1169	174	6
	GO	807	112	6
Netherlands	IO	929	173	13
	K	1348	49	0
	P	4812	346	14
	R	742	92	3
	Z	848	216	12
Australia	B	77	64	23
	C	260	62	22

datasets described in table 2.5. The SparseMSE, dga, LCMCR and Independence estimators are then applied to these datasets, and the point estimates \hat{N} are compared to the ground truth population size N , resulting in the log relative bias, $\log(\hat{N}/N)$.

In table 2.6, we report its empirical mean $\mathbb{E}[\log \hat{N}/N]$, its root mean square error (RMSE) $\mathbb{E}[(\log \hat{N}/N)^2]$, and its median, after removing the outlying results of Netherlands' list K (for which no overlap data is available). Additionally, we report the empirical coverage of 95% confidence intervals. Figure 2.6 shows the estimates and ground truth for every conditioned dataset.

The SparseMSE point estimate appears to perform best, with low mean and

Table 2.6: Summary results of the internal consistency analysis. Best results are bolded in each column.

Estimator	Mean	RMSE	Median	Coverage
dga	-0.34	0.60	-0.22	0.80
Independence	-0.29	0.55	-0.28	0.88
LCMCR	-0.52	0.72	-0.50	0.60
SparseMSE	-0.17	0.63	-0.15	0.90

median log relative bias. Otherwise, the point estimates are all roughly comparable. Regarding confidence intervals, lower bounds are smaller than the ground truth in all cases (excepted with SparseMSE applied to the Netherlands dataset conditioned on list K). It is interesting to note that the Independence model does not perform worse than other approaches.

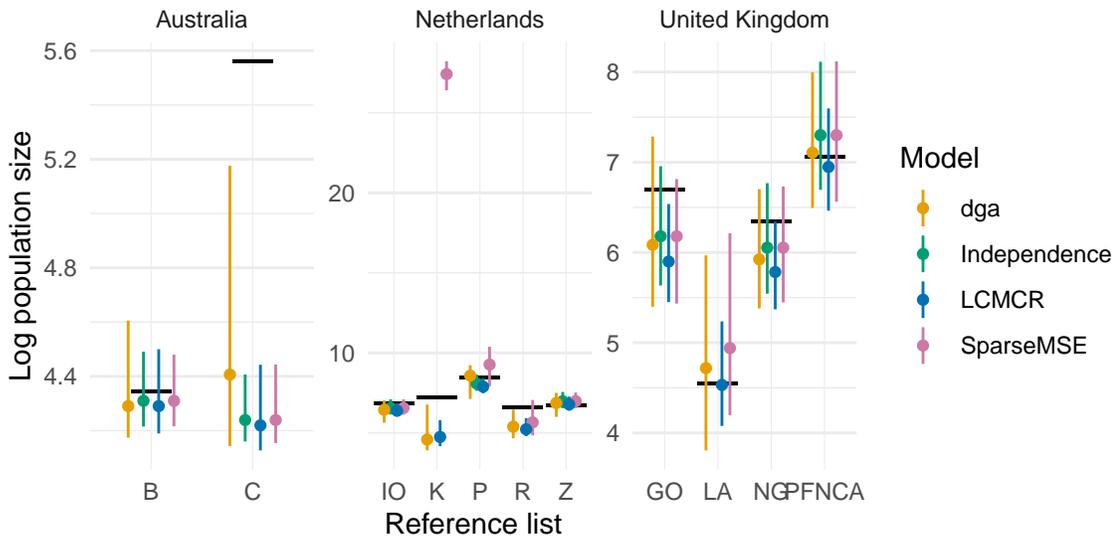


FIGURE 2.6: Results of the internal consistency analysis for every considered dataset and reference list. The black horizontal lines represent ground truth population size. Estimates and 95% confidence intervals are represented by points and vertical lines.

Overall, the results of the internal consistency analysis are highly encouraging. Lower bounds of the confidence intervals are almost always lower than ground truth,

and estimates tend to be close to the ground truth. Coverage of SparseMSE is almost nominal at 90%.

2.4.3 Limitations

The main limitations of the internal consistency analysis are that the conditioned datasets for which ground truth is available are few in number and are not entirely representative of the modern slavery application. There are two main issues here: (1) the conditioned datasets are small and contain little overlap data; and (2) conditioning on a large list necessarily removes from consideration the features of unobserved individuals. That is, regarding point (2), issues of individual heterogeneity and of certain interaction between lists may be removed by conditioning. Our observations regarding the accuracy of estimates on conditioned datasets may therefore not be generalizable to real applications.

Regarding issue (2), section 2.5 next evaluates the bias which can be expected in the presence of individual heterogeneity. That is, we provide a novel characterization of the bias of MSE estimators when underlying assumptions are not satisfied. This characterization is used to compute the bias of estimates under various heterogeneity models. In order to address issue (1), we propose in section 2.6 a visual resampling technique to evaluate the robustness of estimates on practical datasets.

2.5 Bias Under Misspecified Assumptions

Throughout the paper, we have mentioned how population size estimation relies on an untestable assumption of the form **A3**. In this section, we investigate assumption **A3.1** of no full-way interaction term in the log-linear model. We consider the set of estimators which are consistent under **A3.1**, used in modern slavery studies, and we characterize their asymptotic bias when this assumption is misspecified (Theorem 1). In section 2.5.2, we consider the consequences of individual heterogeneity on the

bias of population size estimators. We chose heterogeneity models for their relevance to the modern slavery application, as individual characteristics necessarily affect list inclusion probability. Proposition 2 describes the sign of the bias under a general heterogeneity model. We discuss the case of a Beta heterogeneity model in section 2.5.2.1. Figure 2.7 illustrates the magnitude of the bias under the Beta heterogeneity model as a function of a precision parameter and of the number of lists. The results are summarized in section 2.5.3.

2.5.1 Characterization of the Asymptotic Relative Bias

A standard asymptotic framework for MSE (Chao et al., 2008) considers the large population limit $N \rightarrow \infty$, where the list inclusion patterns $\{W_i\}_{i=1}^\infty$ (see Section 2.2.2) are independent and have distribution (2.3). In addition, the counts $\{n_x\}_{x \neq \mathbf{0}}$ are now a function of N as specified by (2.2). In this context, we can define the consistency property of population size estimators \hat{N} , which are functions of $\{n_x\}_{x \neq \mathbf{0}}$.

Definition 1 (Consistency). A population size estimator \hat{N} is said to be consistent, for the model specified by (2.3), if $\hat{N}/N \rightarrow 1$ almost surely as $N \rightarrow \infty$.

Departure from consistency is quantified by the asymptotic relative bias.

Definition 2 (Asymptotic relative bias). The asymptotic relative bias of a population size estimator, for the model specified by (2.3), is defined as

$$\lim_{N \rightarrow \infty} \frac{\hat{N} - N}{N} \tag{2.10}$$

when this limit is well-defined and almost surely constant.

As noted in Proposition 1, no population size estimator is consistent for all models. Theorem 1 characterizes the asymptotic bias of all estimators which would be consistent under assumption **A3.1**, when in fact this assumption is misspecified.

Note that all convergence statements are understood to happen with probability one (i.e. almost surely).

Theorem 1 (Characterization of the asymptotic relative bias). *Let \hat{N} be any population size estimator which is consistent under assumption **A3.1** of $\gamma = 0$ (no full-way interaction) in the log-linear representation of the model. Then the relative asymptotic bias of \hat{N} exists and is given, when γ is not necessarily equal to zero, by*

$$\lim_{N \rightarrow \infty} \frac{\hat{N} - N}{N} = p_0(e^\gamma - 1). \quad (2.11)$$

See Appendix 2.8.2 for the proof of Theorem 1.

Remark 4 (Lower and upper bound estimators). Theorem 1 shows that population size estimators which would be consistent when $\gamma = 0$ become lower bound estimators when $\gamma < 0$ and upper bound estimators when $\gamma > 0$. That is, when $\gamma < 0$ or $\gamma > 0$, \hat{N} becomes a consistent estimator of a portion or of a multiple of N .

Theorem 1 can be equivalently expressed as providing a first-order approximation to population size estimators.

Corollary 1. *Any population size estimator \hat{N} which is consistent in the absence of full-way interaction term between the lists has the approximation*

$$\hat{N} = \left(1 + \frac{p_0}{1-p_0}e^\gamma + o(1)\right) n_{obs} \quad (2.12)$$

where γ is defined in (2.7) and $o(1)$ is a term which tends to zero as $N \rightarrow \infty$.

Proof. Using the fact that $\lim_{N \rightarrow \infty} n_{obs}/N = 1 - p_0$, we can rearrange (2.11) as

$$\lim_{N \rightarrow \infty} \frac{\hat{N}}{n_{obs}} = 1 + \frac{p_0}{1-p_0}e^\gamma.$$

Equivalently, $\frac{\hat{N}}{n_{obs}} = 1 + \frac{p_0}{1-p_0}e^\gamma + o(1)$ and the result follows directly. \square

2.5.2 Bias in the Presence of Individual Heterogeneity

We now use Theorem 1 to consider the consequences of individual heterogeneity on the bias of estimators which assume no full-way interaction among lists. Recall that these estimators are the ones being used in the context of modern slavery studies. Since individual heterogeneity is to be expected in this application, it is important to evaluate its practical consequences. In this section, we show that individual heterogeneity is incompatible with the assumption of no full-way interaction among lists. Also, we precisely quantify the effect of reasonable heterogeneity models on the bias of estimates.

Assume that each individual $i = 1, 2, \dots, N$ has an individual list appearance probability

$$\lambda_i \sim^{ind.} F \tag{2.13}$$

where F is a distribution supported on $(0, 1]$, and

$$\mathbb{P}(W_i = x \mid \lambda_i) = \lambda_i^{|x|} (1 - \lambda_i)^{L - |x|}, \quad x = (x_1, \dots, x_L) \in \{0, 1\}^L \tag{2.14}$$

where $|x| = \sum_{i=1}^L x_i$. The inclusion patterns W_i are still independent and they are marginally distributed as

$$p_x = \mathbb{P}(W_i = x) = \mathbb{E} \left[\lambda_i^{|x|} (1 - \lambda_i)^{L - |x|} \right]. \tag{2.15}$$

Using the notations of Otis et al. (1978), this is termed an M_h model allowing individual-specific inclusion probabilities.

Proposition 2 illustrates some of the consequences of ignoring heterogeneity for estimators which are consistent under the assumption of no full-way interaction term. In the context of two lists, heterogeneity implies a negative bias. With more than two lists, the bias may be positive or negative. This can be contrasted with the behavior of other classes of estimators. For instance, Horvitz-Thompson type estimators which

wrongly ignore heterogeneity always have a negative bias for any number of lists (Hwang and Huggins, 2005).

Proposition 2. *Let \hat{N} be a population size estimator which is consistent under the assumption $\gamma = 0$ in (2.7). In the context of two lists ($L = 2$) and for the latent heterogeneity model (2.15), necessarily $\lim_{N \rightarrow \infty} \frac{\hat{N} - N}{N} \leq 0$. With three or more lists, the asymptotic bias is positive in some cases and negative in others.*

Proof. From Theorem 1, it suffices to compute γ in the context of the heterogeneity model (2.15). In the two lists setting,

$$\gamma = \log \frac{(\mathbb{E}[\lambda_i(1 - \lambda_i)])^2}{\mathbb{E}[\lambda_i^2] \mathbb{E}[(1 - \lambda_i)^2]} < 0$$

by the Cauchy-Schwartz inequality, and hence \hat{N} is negatively biased, asymptotically. With three and four lists, it is easy to find examples where the bias is positive or negative. \square

2.5.2.1 Beta Heterogeneity Model

In order to make Proposition 2 more concrete, consider the case where

$$\lambda_i \sim^{i.i.d.} \text{Beta}(a, b),$$

and again

$$\mathbb{P}(W_i = x \mid \lambda_i) = \prod_{i=1}^L \lambda_i^{x_i} (1 - \lambda_i)^{1-x_i}, \quad x = (x_1, \dots, x_L) \in \{0, 1\}^L. \quad (2.16)$$

The inclusion patterns W_i are marginally distributed as

$$p_x = \mathbb{P}(W_i = x) \propto \Gamma(a + |x|) \Gamma(b + L - |x|). \quad (2.17)$$

Furthermore,

$$p_0 = \frac{\Gamma(a + b) \Gamma(b + L)}{\Gamma(b) \Gamma(a + b + L)} \approx \left(\frac{b}{a + b} \right)^L \quad (2.18)$$

and

$$\gamma = - \sum_{k=0}^L (-1)^k \binom{L}{k} \log (\Gamma(a+k)\Gamma(b+L-k)).$$

2.5.2.1.1 Two-Lists Beta Model

In the context of two lists, where $L = 2$, Theorem 1 simplifies as

$$\lim_{N \rightarrow \infty} \frac{\hat{N} - N}{N} = -p_{\mathbf{0}} \left(\frac{a+b+1}{(a+1)(b+1)} \right) \leq -p_{\mathbf{0}} \left(\max \left\{ \frac{1}{a+1}, \frac{1}{b+1} \right\} \right) \leq 0.$$

For example, with $a = 1$ and $b = 8$, it follows that 20% of the cases are observed on average. The asymptotic relative bias is $-4/9$ and $\hat{N} \approx \frac{5}{9}N$. As $a \rightarrow 0$, the asymptotic relative bias tends towards -100% .

2.5.2.1.2 Three-Lists Beta Model

In the context of three lists, we obtain

$$\gamma = \log \left(\frac{a(b+1)^2(a+2)}{b(a+1)^2(b+2)} \right).$$

This is positive when $\mathbb{E}[\lambda_i] > 1/2$ and negative when $\mathbb{E}[\lambda_i] < 1/2$; there is a positive bias in the first case and a negative bias in the second. Note, however, that this simple expression for the sign of the bias does not hold outside of the Beta model. In general, the sign of the bias is also linked to higher moments of λ_i .

2.5.2.1.3 Multi-Lists Beta Model

Now consider a Beta model with L lists, where using (2.18) we fix $p_{\mathbf{0}} \approx (b/(a+b))^L = 3/4$ and we let the precision parameter $a+b$ of the list inclusion probabilities λ_i vary. As $a+b$ tends to infinity, individual heterogeneity is reduced, while small positive values of $a+b$ represent high heterogeneity.

Figure 2.7 shows the asymptotic relative bias of population size estimators as a function of the precision $a + b$ of the Beta distribution and of the number of lists L . Note that there is a significant reduction of the relative bias when going from two to three lists. However, differences between using three to six lists are negligible. Furthermore, even for reasonably low heterogeneity levels ($a + b \approx 5$, meaning a standard deviation for the list inclusion probability of about 0.12 when $L = 3$), we find a relative bias of about -50% . This means that estimates will be two times too small.

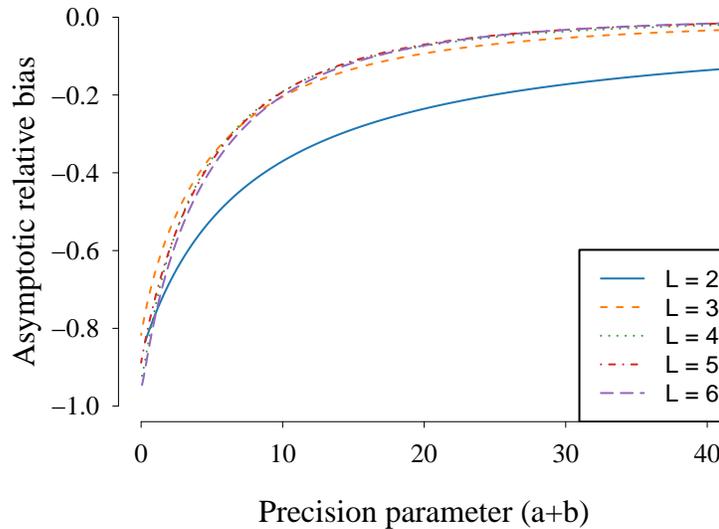


FIGURE 2.7: Asymptotic relative bias of population size estimators (assuming no highest order interaction) as a function of the precision parameter $a + b$ in the Beta model formulation and where we have fixed $p_0 \approx (b/(a + b))^L = 3/4$.

2.5.3 Summary

We have demonstrated how Theorem 1 can be used to evaluate the bias of estimates when the data has characteristics which break the assumption of no full-way interaction term. In particular, individual heterogeneity can result in substantial bias which can be either positive or negative. Furthermore, using more than three lists does

not substantially reduce the consequence of individual heterogeneity under the Beta model which we considered.

The issue of individual heterogeneity has been addressed in some past studies through stratification or modeling. However, other issues remain even if heterogeneity can be accounted for. The most important might be observer effects. For example, an individual observed by one list might cause it to not appear on any other, or there might be systematic referral mechanisms between organizations. Theorem 1 provides an avenue to evaluate the consequences of these data characteristics on the bias (and ultimately accuracy) of estimates.

2.6 Visual Assessment of Robustness

Resampling techniques, including model-based and nonparametric bootstrapping approaches, provide powerful tools for the analysis of estimator properties in application to real data (Efron, 1982). In particular, they may be used to estimate the bias of point estimates and the coverage of confidence intervals. However, such analyses rely on technical assumptions (Hall, 1992). They can be sensitive to modeling assumptions, and theoretical guarantees are only valid in large samples (including the need for large amounts of overlap data). In order to avoid any controversy of the kind found in Whitehead et al. (2019) and Vincent et al. (2020a) regarding the setup of such experiments, we propose a simple visual assessment of estimate robustness and reliability. Our goal is that this visual assessment will be non-controversial and meaningful in practical applications. Our proposal is described in Section 2.6.1. Section 2.6.2 applies our tool to the aforementioned modern slavery datasets.

2.6.1 Visualizing Estimate Trajectories

We propose to consider the series of estimates which would have been obtained if the data had been collected sequentially. That is, we consider series of estimates

obtained as a function of the number of observed individuals. While this depends on the (unknown) order in which individuals have been observed, we may sample the order at random in order to obtain representative samples. The series can also be extended beyond the total size of the reported dataset through resampling.

The behavior of the series may be indicative of convergence towards a stable estimate, or, if it is largely unstable, this can point to potential issues. Its behavior can also be compared to what would be expected under a simple independence model fitted to the data, or under a more complex model fitted to the data. Differences between the behaviors of the series would then indicate a lack of fit of these simple or more complex models.

To be more precise, consider a dataset $\mathcal{D} = \{W_i\}_{i=1}^n$ of n observations, where each $W_i \in \{0, 1\}^L$ is an observed list inclusion pattern. From this dataset, we construct a series $\{Z_i\}_{i=1}^n$ which represents a hypothetical ordering of the observations in \mathcal{D} . This is obtained by choosing a permutation σ of $\{1, 2, \dots, n\}$ at random and setting $Z_i = W_{\sigma(i)}$. Furthermore, the series Z_i is extended to $2n \geq i > n$ by choosing a second random permutation π and setting $Z_{n+i} = W_{\pi(i)}$ for $1 \leq i \leq n$. This series $\{Z_i\}$ represents an hypothetical sample path of list inclusion pattern. The corresponding population size estimates, each computed using the first n_{obs} data points $\{Z_i\}_{i=1}^{n_{\text{obs}}}$, are the estimate trajectories which we focus on.

Figure 2.8 shows such trajectories of dga estimates on data from an independence model fitted to the United Kingdom dataset. That is, we have simulated a single dataset from the independence model and then applied our proposed procedure to it which provides different estimate trajectories for this data. Panel **A** shows a single trajectory with point estimates and confidence intervals. Panel **B** shows point estimate trajectories corresponding to 50 random orderings. The horizontal dotted line represents ground truth. The behaviour of these trajectories can be considered a best case scenario, given that the data came from a simple independence model.

This can be compared with the application to real data in the following section.

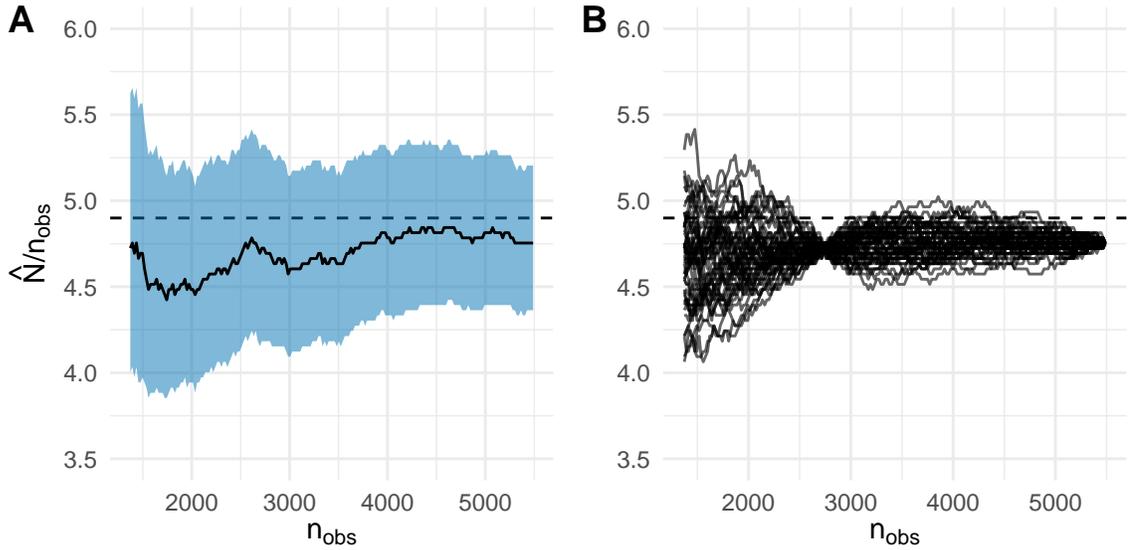


FIGURE 2.8: Trajectories of dga estimates (as the ratio of estimated population size to number of observations) on data simulated from and independence model fit to the United Kingdom dataset. The horizontal line represents ground truth population size. Panel **A** shows a single trajectory together with the 95% credible intervals. Panel **B** shows 50 random trajectories for the same dataset.

2.6.2 Application to Real Data

We now present the result of our visualization in application to the United Kingdom and Netherlands datasets. We focus on these two datasets because they are the largest and they are the ones for which sensitivity to individual observations is most noticeable.

Figure 2.9 shows trajectories of estimates in application to the United Kingdom dataset. Figure 2.10 shows trajectories of dga and SparseMSE estimates in application to the Netherlands dataset. LCMCR and independence estimate trajectories have been omitted from figure 2.10 since, like in the case of the United Kingdom data, they did not showcase high sensitivities. The thin vertical lines indicate the number of observations at which the estimate trajectories coincides with real data estimates.

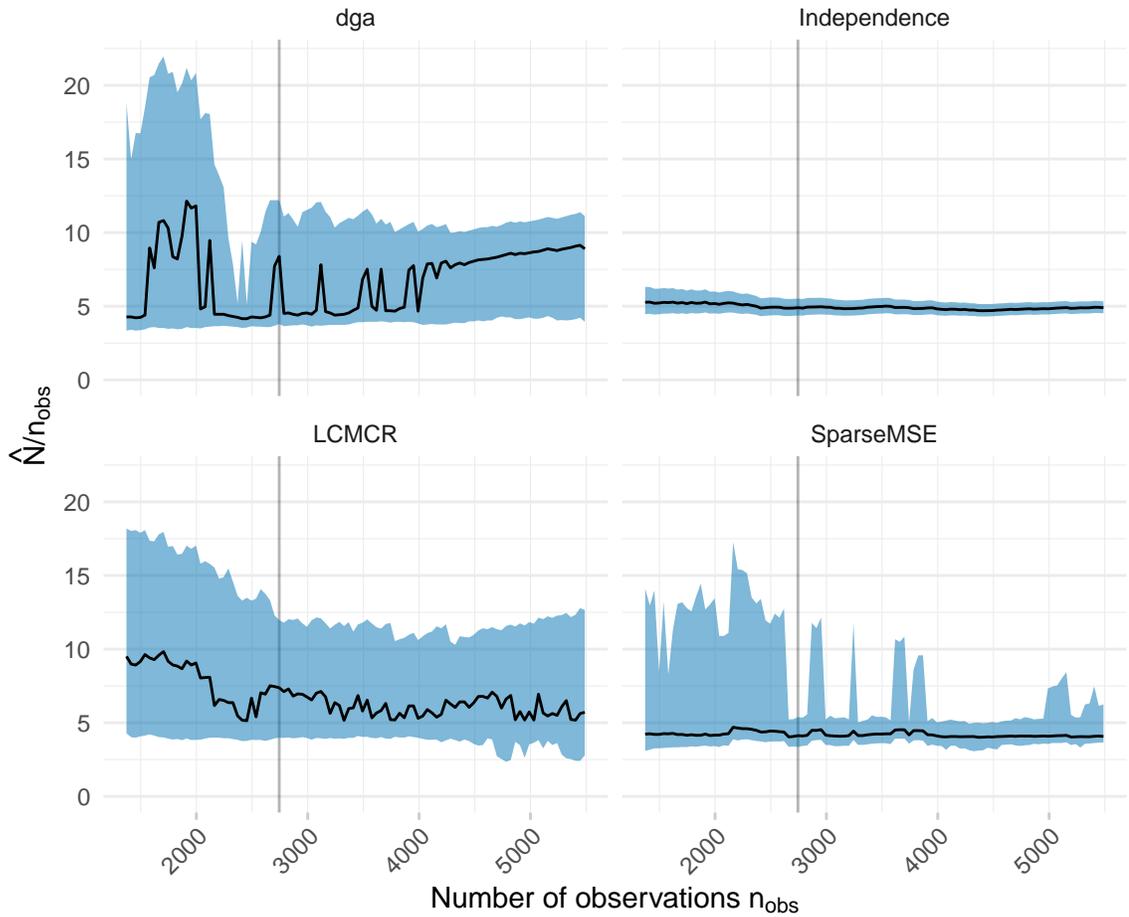


FIGURE 2.9: Visualization of estimate trajectories (as the ratio of estimated population size to number of observations) for the United Kingdom dataset. The horizontal line represents ground truth population size. The thin vertical lines indicate the number of observations in the United Kingdom dataset, at which the trajectory estimates coincide with real data estimates.

Looking at Figure 2.9, the Independence and LCMCR estimates appear quite stable on the United Kingdom dataset. On the other hand, the dga and SparseMSE estimates are much more sensitive to individual observations. Regarding dga estimates, while 95% credible intervals are relatively stable, the point estimates (posterior median) can significantly change as the result of observing only a few additional cases. Regarding SparseMSE estimates, the narrow confidence interval obtained on the United Kingdom

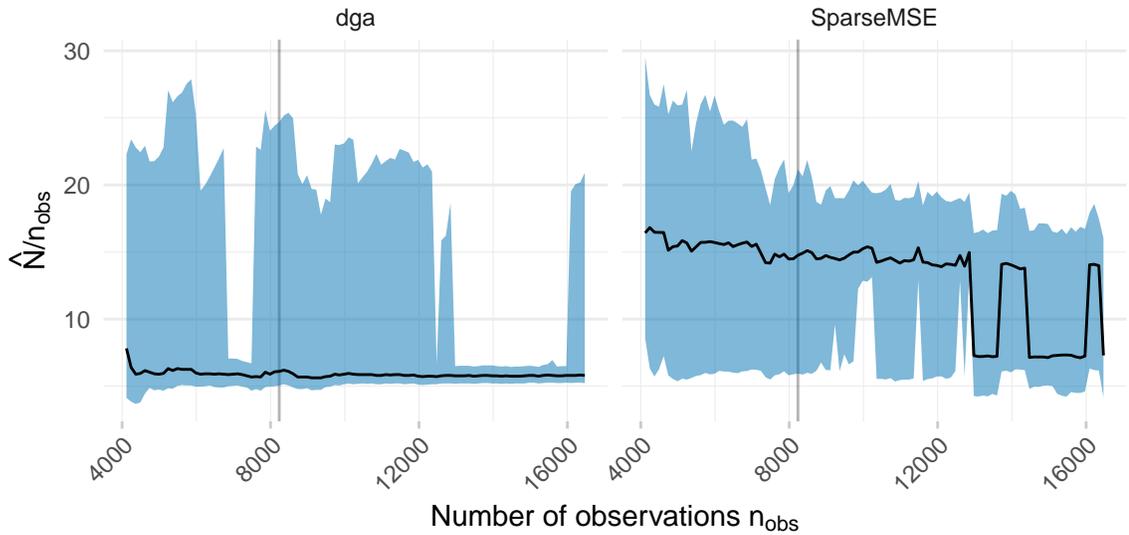


FIGURE 2.10: Visualization of estimate trajectories (as the ratio of estimated population size to number of observations) for the Netherlands dataset and for the dga and SparseMSE estimators. The thin vertical lines indicate the number of observations in the United Kingdom dataset, at which the trajectory estimates coincide with real data estimates.

dataset appears to be a fluke. With only a few less or a few more observations, the confidence interval becomes much larger, to between around 4 and 12 times the number of observed cases. This suggests that bootstrap confidence intervals used by SparseMSE may sometimes fail to properly account for uncertainty.

In the case of the Netherlands dataset shown in Figure 2.10, the behavior of the dga and sparseMSE estimates is quite different. The 95% credible regions of dga now fluctuate much more, while SparseMSE point estimates are less stable in larger samples.

In practice, the estimate trajectories can help diagnose lack of robustness. In the case of the SparseMSE estimate for the United Kingdom dataset, for instance, Figure 2.9 shows strong sensitivities and confidence intervals which are too narrow. This should be addressed by considering a broader range of plausible estimates.

2.7 Discussion

MSE has unique potential in helping assess the true scale of modern slavery. However, long-standing controversy in the literature have come in the way of the broader implementation of MSE methodology. As such, we address three major issues debated recently, namely, statistical aspects of MSE assumptions, robustness, and accuracy. First, we review the current state of the MSE literature, commonly used methods, and all publicly available modern slavery datasets. Next, we provide a reproducible analyses evaluating the accuracy of estimates, the consequences of MSE assumptions, and the robustness of estimates to small changes in data. Specifically, we utilize the internal consistency approach of Hook et al. (2012) to evaluate the accuracy of estimates when ground truth is available (section 2.4). Then, we assess the consequences of MSE assumptions through a novel characterization of large sample bias (section 2.5). Finally, we propose a visual assessment of reliability and robustness (section 2.6). Our work highlights important practical and methodological challenges with MSE, which we summarize below. In addition, we comment on future methodological research.

2.7.0.0.1 Practical Challenges and Recommendations Our work highlights some important statistical challenges practitioners face when using MSE to quantify modern slavery. First, we have shown in section 2.3.4 how estimates can be highly sensitive to the choice of tuning parameters. Due to this, we recommend for sensitivity analyses to be conducted, evaluated, and reported in all applications. Second, we have shown (section 2.6) that estimates can be highly sensitive to individual data points. Our proposed visualization of estimate trajectories can be used by practitioners as a diagnostic tool moving forward in analyses. Third, we have shown (section 2.5) that assumptions underlying MSE are highly influential. In particular, we have quantified

how reasonable individual heterogeneity models can affect the bias of estimates from models used in practice. Thus, we recommend that users report, discuss, and justify underlying assumptions in their work moving forward. Ideally, enough information about data collection should be provided for these assumptions to be scrutinized. To summarize, this provides a set of simple guidelines and recommendations that can be used in MSE studies, complementing other recommendations that have been put forth (Hook and Regal, 1999).

2.7.0.0.2 Future Methodological Research

The statistical challenges which we have highlighted point at a problem of under-specification (D’Amour et al., 2020). In our context, this means that many seemingly reasonable approaches to MSE can give different results. There are two main reasons for this. First, there is the choice of underlying assumption of the form **A3** which we have discussed and shown to be highly influential. Aleshin-Guendel et al. (2021) has provided first steps towards emphasizing the role of these assumptions, but further work is needed to assess their suitability in real applications and to propose meaningful alternatives. Second, even if an assumption of the form **A3** can be justified, we have shown how estimates can be highly sensitive to tuning parameters and individual data points. That is, small errors or changes in data and arbitrary choices in model fitting procedures can lead to vastly different estimates. Methodology accounting for noise in the data, record linkage errors, and other sources of uncertainty, are needed to address these issues.

2.7.0.0.3 Software

Code and data for all analyses presented in this paper are available at <http://github.com/OlivierBinette/MSETools>

2.7.0.0.4 Appendix

The appendix contains summaries of the datasets considered in this paper and proofs which were omitted from the main text.

2.8 Appendix

2.8.1 Datasets Summary

United Kingdom Silverman (2014); Bales et al. (2015) considered data on potential victims of modern slavery collected as part of the National Crime Agency Strategic Assessment (United Kingdom Human Trafficking Centre, 2014). This data identifies potential victims reported by six different sources of information in 2013: the police force (PF), the National Crime Agency (NCA), local authorities (LA), governmental organizations (GO), non-governmental organizations (NG), and the general public (GP). A total of 2744 cases are recorded, including 221 cases which appeared on more than one list. No cases appear on both the LA and GP lists, nor on both the LA and NCA lists. This may be due to the small size of the LA list (94 cases), to the way that the lists were constructed as part of the Strategic Assessment, or due to other unknown factors. No covariate information, such as the type of exploitation, country of origin, or demographic variables, is available for the potential victims. The mechanisms through which cases came to be observed on only one list or on more than one lists, such as the referral pathways between organizations, are not described. Using this data, Silverman (2014); Bales et al. (2015) inferred a total of between 10,000 and 13,000 potential victims in the UK in 2013.

New Orleans Bales et al. (2019) carried out a similar MSE study in the Greater New Orleans region of the United States. Eight (anonymous) organizations reported together a total of 185 “confirmed cases of human trafficking” for 2016,

including 14 victims who appeared on more than one list. The authors noted challenges associated with this data, notably the definitional ambiguity of human trafficking cases and the differing goals of some organizations. One organization worked only with victims of labor trafficking, one only with victims of labor trafficking who were also victims of sex trafficking, another only reported victims of both sex trafficking and labor trafficking, and five only reported victims of sex trafficking. No covariate information is publicly available nor used in this study. Using this data, the authors inferred between around 650 and 1700 victims in 2016.

Netherlands van Dijk et al. (2017) estimated the number of trafficked victims in the Netherlands disaggregated by sex, age, form of exploitation, nationality, and year between 2010 and 2015. This used data from a state-sponsored NGO, CoMensha, to which presumed cases of human trafficking are reported by the police and other organizations. The sources of information were grouped into six lists, the border police (K), the inspectorate (I), residential centers/shelters (O), national police (P), regional coordinators (R) and others (Z). A total of with 8,234 cases were observed, including 432 cases which appeared on more than one list. Only the aggregated data, with no covariate information, is publicly available. The authors noted challenges with this data, notably the definitional ambiguity of reported cases. In particular, the border police reported cases which do not correspond to the international definition of human trafficking. This lead the authors to produce estimates for both the inclusion and exclusion of this list. The authors also noted that the inclusion of covariates resulted in a significant change in estimated population size. Using covariates and all lists, they estimated around 42,000 victims. Using the same approach but ignoring covariates, they obtained an estimate of around 58,000 victims.

Western U.S. Farrel et al. (2019) investigated the reporting and data collection mechanisms in place at law enforcement organizations regarding cases of human trafficking. They also carried out MSE studies at two locations, including at the “Western site” for which aggregated data is publicly available. This data records a total of 345 individuals which appeared in four lists, including a law enforcement list and three community service provider lists. A total of 23 individuals were observed in more than one list. The authors used covariate information with multiple imputation for missing data in order to obtain an estimate of between around 1,600 and 3,600 victims in the Western site in 2016.

Australia Lyneham et al. (2019) reproduced the study of Silverman (2014) with data collected by the Australian Federal Police. The dataset which appears in Lyneham et al. (2019) and which we have use consists of four lists and a total of 414 cases of victims observed between 2015-2016 and 2016-2017. The authors estimated between 900 to 1,500 total victims for this time period.

Other studies Other MSE studies for the quantification of modern slavery have been carried out in Serbia, Ireland, Romania and Slovakia (UNODC, 2018a,b,c; Vincent et al., 2020a). Data is not publicly available in these cases.

2.8.2 Proof of Theorem 1

The proof of the Theorem is separated in four parts. First we set up notation and the background of the problem. Steps 1-3 then provide the main ingredients necessary to establish (2.11). While the proof is not complex, it does require a certain level of detail. Note that, throughout, convergence is meant to be in the almost sure sense.

Setup

Let us make explicit the dependency of the counts n_x on N by writing

$$n_x^{(N)} = \#\{i \leq N : W_i = x\}, \quad x \in \{0, 1\}^L \setminus \{\mathbf{0}\}, n^{(N)} = \sum_{x \neq \mathbf{0}} n_x^{(N)}, \quad (2.19)$$

where W_i , $i = 1, 2, 3, \dots$, is an independent sequence defined as in (2.3).

The count process $(n_x^{(N)})_{x \neq \mathbf{0}}$, for $N = 1, 2, 3, \dots$, is a Markov chain with distribution defined through the following:

1. given $(n_x^{(N)})_{x \neq \mathbf{0}}$, with probability p_0 we have $(n_x^{(N+1)})_{x \neq \mathbf{0}} = (n_x^{(N)})_{x \neq \mathbf{0}}$;
2. otherwise, for $x \in \{0, 1\}^L \setminus \{\mathbf{0}\}$ distributed with probability mass function q_x , we have $n_x^{(N+1)} = n_x^{(N)} + 1$ and $n_{x'}^{(N+1)} = n_{x'}^{(N)}$ for every $x' \neq x$.

Now recall that if $\gamma = 0$, then for any sequence $(n_x^{(N)})_{x \neq \mathbf{0}}$ distributed as above we have

$$\hat{N}((n_x^{(N)})_{x \neq \mathbf{0}}) / N \rightarrow 1 \quad (2.20)$$

(almost surely) as $N \rightarrow \infty$, by assumption and definition of consistency.

Here, since we do not assume that $\gamma = 0$, our argument instead relies on the fact that there exists a random subsequence N_k , $k = 1, 2, 3, \dots$, such that $(n_x^{(N_k)})_{x \neq \mathbf{0}}$, $k = 1, 2, 3, \dots$ corresponds to the count process of a model with no full-way interaction and therefore

$$\hat{N}((n_x^{(N_k)})_{x \neq \mathbf{0}}) / k \rightarrow 1 \quad (2.21)$$

as $k \rightarrow \infty$. This is shown in Step 1 below. In Step 2, we show that the sequence N_k satisfies $N_k/k \rightarrow (1 - p'_0)/(1 - p_0)$ almost surely as $k \rightarrow \infty$, allowing us to compute the limit of the ratio $\hat{N}((n_x^{(N_k)})_{x \neq \mathbf{0}}) / N_k$. Step 3 shows that the relative asymptotic bias of \hat{N} exists and we then easily deduce its form.

Step 1

Here we define a random sequence N_k , $k = 1, 2, 3, \dots$, such that $\hat{N}((n_x^{(N_k)})_{x \neq \mathbf{0}})/k \rightarrow 1$ as $k \rightarrow \infty$. Let $p'_0 > 0$ be defined as

$$\frac{p'_0}{1 - p'_0} = e^\gamma \frac{p_0}{1 - p_0}; \quad p'_0 = \frac{p_0 e^\gamma}{1 - p_0 + p_0 e^\gamma} \quad (2.22)$$

and for $x \neq \mathbf{0}$ let $p'_x = (1 - p'_0)q_x$. The probabilities p'_x , $x \in \{0, 1\}^L$, define an alternative model to (2.3) for which

$$q'_x = \frac{p'_x}{(1 - p'_0)} = q_x \quad \text{and} \quad \gamma' = \log \left(\frac{1 - p'_0}{p'_0} \frac{q'_{\text{odd}}}{q'_{\text{even}}} \right) = 0. \quad (2.23)$$

Now let us define the random sequence N_k as follows as a function of $\{(n_x^{(N)})_{x \neq \mathbf{0}}\}_{N=1}^\infty$. Let $0 = s_0 < s_1 < s_2 < s_3 < \dots$ be the sequence of integers such that s_i is the smallest integer satisfying $n^{(s_i)} = i$. Let t_1, t_2, t_3, \dots be an i.i.d. sequence of geometric random variables with mean $1/(1 - p'_0)$, let $T_j = \sum_{i \leq j} t_i$, and let

$$N_k = \sum_{i: T_i \leq k} (s_i - s_{i-1}). \quad (2.24)$$

That is, N_k is such that $N_k = s_j$ when $T_j \leq k < T_{j+1} - 1$.

By construction, the distribution of $(n_x^{(N_{k+1})})_{x \neq \mathbf{0}}$ given $(n_x^{(N_k)})_{x \neq \mathbf{0}}$ is the following:

1. $n^{(N_{k+1})} = n^{(N_k)}$ with probability p'_0 ;
2. otherwise $n_x^{(N_{k+1})} = n_x^{(N_k)} + 1$ for some x distributed following q'_x and $n_{x'}^{(N_{k+1})} = n_{x'}^{(N_k)}$ for $x' \neq x$.

This is exactly the distribution of a count process obtained from the model with probabilities p'_x defined in (2.22) and (2.23). These probabilities satisfy the no full-way

interaction assumption **A3.1**, and it follows from the consistency of the estimator \hat{N} that

$$\hat{N}((n_x^{(N_k)})_{x \neq \mathbf{0}})/k \rightarrow 1 \quad (2.25)$$

as $k \rightarrow \infty$.

Step 2

This step shows that $N_k/k \rightarrow (1 - p'_0)/(1 - p_0)$ as $k \rightarrow \infty$. Note that the variables $s_i - s_{i-1}$ are independent with a geometric distribution of mean $1/(1 - p_0)$ and they are independent from the variables T_j . By the strong law of large numbers, we have

$$N_{T_j}/j = \frac{1}{j} \sum_{i=1}^j (s_i - s_{i-1}) \rightarrow (1 - p_0)^{-1} \quad (2.26)$$

and

$$T_j/j = \frac{1}{j} \sum_{i=1}^j t_i \rightarrow (1 - p'_0)^{-1} \quad (2.27)$$

as $j \rightarrow \infty$. Hence dividing the two we obtain $N_{T_j}/T_j \rightarrow (1 - p'_0)/(1 - p_0)$ as $j \rightarrow \infty$. Now for $j = j(k)$ such that $T_j \leq k < T_{j+1}$, we have $N_k = N_{T_j}$ and

$$N_k/k = \frac{N_{T_j} T_j}{T_j k} \rightarrow (1 - p'_0)/(1 - p_0) \quad (2.28)$$

follows from the fact that $T_j/k \rightarrow 1$ as $k \rightarrow \infty$.

Step 3

Combining steps 1 and 2, we obtain that $\hat{N}((n_x^{(N_k)})_{x \neq \mathbf{0}})/N_k \rightarrow (1 - p_0)/(1 - p'_0)$ as $k \rightarrow \infty$. It can easily be verified that $(n_x^{(N)})_{x \neq \mathbf{0}}$ is also a subsequence of $(n_x^{(N_k)})_{x \neq \mathbf{0}}$, and therefore \hat{N}/N converges with

$$\hat{N}/N = \hat{N}((n_x^{(N)})_{x \neq \mathbf{0}})/N \rightarrow (1 - p_0)/(1 - p'_0)$$

Using expression (2.22) for p'_0 and simplifying, we obtain (2.11). This concludes the proof.

2.8.3 Proof of Proposition 1

Suppose there exists a consistent population size estimator \hat{N} for a model for which no function f satisfies the condition of Proposition 1. That is, there exists two sets of probabilities p_x and \tilde{p}_x , $x \in \{0, 1\}^L$, such that $p_{\mathbf{0}} \neq \tilde{p}_{\mathbf{0}}$ while $q_x = p_x/(1 - p_{\mathbf{0}}) = \tilde{p}_x/(1 - \tilde{p}_{\mathbf{0}}) = \tilde{q}_x$ for $x \neq \mathbf{0}$.

Continuing with similar notations as in the proof of Theorem 1, let $n_x^{(N)}$ and $\tilde{n}_x^{(N)}$, $x \neq \mathbf{0}$, be the two sequences of observed counts corresponding to the model probabilities p_x and \tilde{p}_x , respectively. Let $n^{(N)} = \sum_{x \neq \mathbf{0}} n_x^{(N)}$, $\tilde{n}^{(N)} = \sum_{x \neq \mathbf{0}} \tilde{n}_x^{(N)}$, and let s_k and \tilde{s}_k be sequences of the smallest integers satisfying $n^{(s_k)} = k$ and $\tilde{n}^{(\tilde{s}_k)} = k$.

Since $\hat{N}((n_x)_{x \neq \mathbf{0}}^{(s_k)})$ is a subsequence of $\hat{N}((n_x)_{x \neq \mathbf{0}}^{(N)})$ and by consistency of \hat{N} , we have that $\hat{N}((n_x)_{x \neq \mathbf{0}}^{s_k})/s_k \rightarrow 1$ almost surely as $k \rightarrow \infty$. For the same reasons, $\hat{N}((\tilde{n}_x)_{x \neq \mathbf{0}}^{(\tilde{s}_k)})/\tilde{s}_k \rightarrow 1$ almost surely as $k \rightarrow \infty$.

Now notice that the two sequences $(n_x)_{x \neq \mathbf{0}}^{s_k}$ and $(\tilde{n}_x)_{x \neq \mathbf{0}}^{\tilde{s}_k}$ have exactly the same distribution since $q_x = \tilde{q}_x$, $x \neq \mathbf{0}$. Therefore $\hat{N}((n_x)_{x \neq \mathbf{0}}^{(s_k)})/\hat{N}((\tilde{n}_x)_{x \neq \mathbf{0}}^{(\tilde{s}_k)}) \rightarrow 1$ as $k \rightarrow \infty$.

It follows from the above that $s_k/\tilde{s}_k \rightarrow 1$. However, this can only happen if $p_{\mathbf{0}} = \tilde{p}_{\mathbf{0}}$, which is not the case here. Indeed, the increments $s_{k+1} - s_k$ are independent with geometric distribution of mean $(1 - p_{\mathbf{0}})^{-1}$, and by the law of large numbers it follows that $s_k/k \rightarrow (1 - p_{\mathbf{0}})^{-1}$. Similarly, $\tilde{s}_k/k \rightarrow (1 - \tilde{p}_{\mathbf{0}})^{-1}$, from which it follows that $s_k/\tilde{s}_k \rightarrow (1 - \tilde{p}_{\mathbf{0}})/(1 - p_{\mathbf{0}}) \neq 1$.

3. Estimating the Performance of Entity Resolution Algorithms: Lessons Learned Through PatentsView.org

This chapter reproduces Binette et al. (2023). Motivated by PatentsView.org, a U.S. Patents and Trademarks Office patent data exploration tool that disambiguates patent inventors using an entity resolution algorithm, the chapter introduces a novel evaluation methodology for entity resolution algorithms. We provide a data collection methodology and tailored performance estimators that account for sampling biases. Our approach is simple, practical and principled – key characteristics that allow us to paint the first representative picture of PatentsView’s disambiguation performance. This approach is used to inform PatentsView’s users of the reliability of the data and to allow the comparison of competing disambiguation algorithms.

3.1 Introduction

Entity resolution (also called record linkage, deduplication, or disambiguation) is the task of identifying records in a database that refer to the same entity. An entity may be a person, a company, an object or an event. Records are assumed to contain partially identifying information about these entities. When there is no unique identifier (such as a social security number) available for all records, entity resolution becomes a complex problem which requires sophisticated algorithmic solutions (Herzog et al., 2007; Christen and Christen, 2012; Dong and Srivastava, 2015; Ilyas and Chu, 2019;

Christophides et al., 2021; Christen, 2019; Papadakis et al., 2021; Binette and Steorts, 2022a).

For instance, the U.S. Patents and Trademarks Office (USPTO) makes available patent data dating back to 1790 (digitized full-text data is available from 1976). However, there is no standard for uniquely identifying inventors on patent applications. The result is a set of ambiguous mentions of inventors, where a single person’s name may be spelled in different ways on two applications and where two different inventors with the same name may be difficult to distinguish. Inventor mobility further complicates the use of contextual information for disambiguation.

The problem of disambiguating inventor mentions has attracted much interest in the field of economics, computer science, and statistics. Following seminal works (Trajtenberg and Shiff, 2008; Ferreira et al., 2012; Ventura et al., 2013; Li et al., 2014), a disambiguation competition was held in 2015 leading to the disambiguation system currently used by PatentsView.org within the USPTO. Since then, disambiguation of U.S. patents data has been of continued research interest (Ventura et al., 2015; Kim et al., 2016b; Yang et al., 2017; Morrison et al., 2017; Müller, 2017; Traylor et al., 2017; Balsmeier et al., 2018; Tam et al., 2019; Monath et al., 2019; Doherr, 2021a). To provide more details, PatentsView.org is a patent data platform maintained by the Office of Chief Economist at the USPTO and the American Institutes for Research (AIR). Through its data visualizations, search tools, data products, and Application Programming Interface (API), PatentsView now serves a wide variety of audience including students, educators, researchers, policymakers, small business owners, and the public (Toole et al., 2021). Since the first release of the results from the disambiguation competition held in 2015, the disambiguated inventor data continues to be one of PatentsView’s most popular data product.

Unfortunately, evaluating the performance of entity resolution systems is a difficult problem. In the case of PatentsView’s disambiguation system, no principled evaluation

methodology is available to measure performance, to inform users of the reliability of the data, and to support methodological research to improve upon PatentsView’s disambiguation algorithms. The state-of-the-art in entity resolution evaluation, namely computing performance evaluation metrics (precision, recall, etc.) on benchmark datasets, leads to misleading and highly biased performance metrics as shown in section 3.1.1. This is concerning given the many scientific uses of PatentsView’s data: prior to June 2021, 179 research studies cited PatentsView as a data source, including around 25% from the field of economics (Toole et al., 2021). A common theme of research is the study of the relationship between public policy, inventor mobility, and inventor demographics, on innovation and patenting. This requires accurate inventor disambiguation to track inventors and entities through the breadth of patent data.

Our paper therefore addresses the problem of evaluating the accuracy of PatentsView’s disambiguation, for the purpose of informing users of the reliability of the data and in order to support methodological research to improve upon PatentsView’s disambiguation algorithm. We propose novel evaluation methodology that is principled and cost-effective, and we demonstrate its effectiveness to evaluate PatentsView’s disambiguation. We expand on the challenges of evaluation, past work, and our contributions, in section 3.1.1. Below, we review terminology used throughout.

3.1.0.1 Terminology

We consider a database of records, where each record represents a mention to a given inventor (e.g., the first inventor of Patent number 12345). In this context, the records are also referred to as inventor mentions. The goal of entity resolution is to cluster inventor mentions according to the entity (real-world inventor) to which they refer. Clusterings obtained from algorithms are referred to as predicted clusters or predicted disambiguations, whereas the (unknown) clustering corresponding to the true set of inventors is referred to as the ground truth. Two inventor mentions are said to match

or to be a true match if they refer to the same inventor. If two inventors are in the same predicted cluster, then they are a link, or a predicted match. The proportion of true matches among all predicted matches is called the pairwise precision, while the proportion of predicted matches among all true matches is called the pairwise recall.

3.1.1 The Evaluation Problem

The entity resolution evaluation problem is to extrapolate from observed performance in small samples to real performance in a database with millions of records. Wang et al. (2022) refer to this as bridging the reality-ideality gap in entity resolution, where high performance on benchmark datasets often does not translate into the real world. Here, performance may be defined as any combination of commonly used evaluation metrics for entity resolution, such as precision and recall, cluster homogeneity and completeness, rand index, or generalized merge distance (Maidasani et al., 2012). These metrics can be computed on benchmark datasets for which we have a ground truth disambiguation. However, the key evaluation problem is to obtain estimates that are representative of performance on the full data, for which no ground truth disambiguation is available. This is challenging for the following reasons.

First, entity resolution problems do not scale linearly. While it may be easy to disambiguate a small dataset, the opportunity for errors grows quadratically in the number of records. As such, we may observe good performance of an algorithm on a small benchmark dataset, while the true performance on the entire dataset may be something else entirely. This particular effect of dataset size in entity resolution is explored in Draibach and Naumann (2013) in the context of choosing similarity thresholds. This is a problem that PatentsView.org currently faces. Despite encouraging performance evaluation metrics on benchmark datasets, with nearly perfect precision and recall reported in the latest methodological report (Monath et al., 2021), the data science team at AIR observes lower real-world accuracy. This

phenomenon is illustrated in example 1 below.

A second problem is large class imbalance in entity resolution (Marchant and Rubinstein, 2017). Viewing entity resolution as a classification problem, the task is to classify record pairs as being a match or non-match. However, among all pairs of records, only a small fraction (usually much less than a fraction of a percent) refer to the same entity. The vast majority of record pairs are not a match. This makes it difficult to evaluate performance through random sampling of record pairs.

A third problem is the multiplicity of sampling mechanisms used to obtain benchmark datasets. To construct hand-disambiguated datasets, blocks, entity clusters, or predicted clusters may be sampled with various probability weights. These sampling approaches must be accounted for in order to obtain representative performance estimates (Fuller, 2011).

Our approach, detailed in sections 3.1.1.3 and 3.2.3, addresses these challenges by putting forward novel cluster-based expressions for performance metrics that reflect various sampling schemes. Each of these representations immediately suggests simple estimators that properly account for the above issues.

Example 1 (Bias of precision computed on benchmark datasets). To exemplify the problem with the trivial use of performance evaluation metrics on benchmark datasets, we carried out a toy experiment that is described in detail in appendix 3.5.1. In short, we evaluated a disambiguation algorithm by sampling ground truth clusters and computing pairwise precision on this set of sampled clusters. This is analogous to the way that many real-world benchmark datasets are obtained and typically used. In this experiment, we know that the disambiguation algorithm has a precision of 52% for the entire dataset.

In panel **A** of figure 3.1, we see the distribution of precision estimates versus the true precision of 52% shown as a dotted vertical line. Precision estimates are usually

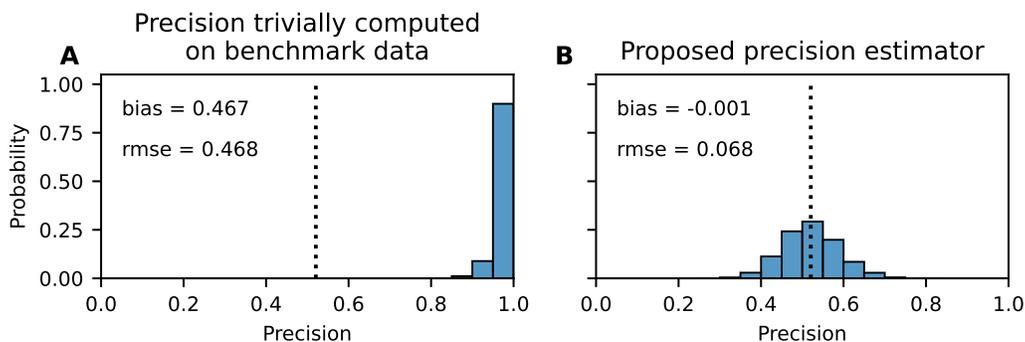


FIGURE 3.1: Distribution of precision estimates versus the true precision of 52% (shown as a dotted vertical line). Panel **A** shows the trivial precision estimates computed for sampled records. Panel **B** shows our proposed precision estimates which accounts for the sampling mechanism. Sample bias and root mean squared error (rmse) are reported in each figure.

very close to 100% and always higher than 80%, despite the truth being a precision of only 52%. In contrast, panel **B** shows the distribution of our proposed precision estimator which is nearly unbiased. Both precision estimators rely on exactly the same data. They only differ in how they account for the underlying sampling process and the extrapolation from small benchmark datasets to the full data.

The same phenomenon can be observed in PatentsView’s data, where naive precision is nearly 1 on all benchmark datasets. In our simulation studies (see Figure 3.3 for instance), naive precision estimates are always nearly 1, despite the true precision ranging from 60% to 90%.

The reason why naive performance estimation performs disastrously is that it is much more easy to disambiguate a small benchmark dataset than a large population with millions of records. Indeed, as a dataset grows, opportunity for erroneous links grows quadratically. False links between similarly named inventor, which are common in the full data, disappear when the benchmark dataset only contains a random sample of inventors. Our performance estimators extrapolate from performance observed on a small benchmark to true performance on the full data.

3.1.1.1 Why Bother With Evaluation?

There are two main uses for accurate and statistically rigorous evaluation methodology.

The first is model selection and comparison. PatentsView.org continually works at improving disambiguation methodology. This requires choosing between alternative methods and evaluating the results of methodological experiments. Without sound evaluation methodology, decisions regarding the disambiguation algorithm may not align with real-world use and real-world performance. Notably, for a performance metric such as the pairwise f-score, one algorithm may perform better than another on a small benchmark dataset, while the opposite may hold true for performance on the entire data. This problem arises with typical benchmark datasets obtained from randomly sampling blocks or randomly sampling clusters (see section 3.2.3 for a definition of different sampling mechanisms).

The second is adequate use of disambiguated data. PatentsView.org’s disambiguation results have been used in numerous scientific studies (Toole et al., 2021). For example, Choudhury and Kim (2019) studied the effect of skilled worker immigration on patenting at U.S. companies and institutions, using PatentsView’s inventor disambiguation to track individual immigrant inventors across time and location. These studies make assumptions about the reliability of the data that need to be validated and upheld. In short, users of disambiguated data need to understand its reliability in order to make scientifically appropriate use of it. Evaluation aims to provide this rigorous reliability information.

3.1.1.2 Past Work

Much of the past literature has focused on defining and using relevant clustering evaluation metrics. The topic of estimating performance from samples has received much less attention, usually focusing on importance sampling estimators based on record pairs. We review the contributions to these two main topics below.

3.1.1.2.1 Metrics

Pairwise precision and recall metrics were first reported in Newcombe et al. (1959), with Bilenko and Mooney (2003) and Christen and Goiser (2007) emphasizing the importance of precision-recall curves for algorithm evaluation. However, there are issues with the use of pairwise precision and recall in entity resolution applications, such as the large relative importance of large clusters. As such, other clustering metrics have been proposed, including cluster precision and recall, cluster homogeneity and completeness, the B^3 metric (Bagga and Baldwin, 1998), and generalized merge distances (Michelson and Macskassy, 2009; Menestrina et al., 2010; Maldasani et al., 2012; Barnes, 2015). Important practical issues regarding the use of aggregate metrics are discussed in Hand and Christen (2018).

While our work focuses on estimating pairwise precision and recall, the general approach also applies to any cluster-based performance evaluation metrics, such as those described in Michelson and Macskassy (2009).

3.1.1.2.2 Estimation

Regarding the reliable estimation of performance metrics, Belin and Rubin (1995) first proposed a semi-supervised approach to calibrating error rates when using a Fellegi-Sunter model (Fellegi and Sunter, 1969). Marchant and Rubinstein (2017) proposed an adaptive importance sampling estimator to estimate precision and recall from sampled record pairs. Other approaches to the estimation of performance metrics are model based, where estimated precision and recall can be obtained from predicted match probabilities between record pairs (Enamorado et al., 2019a). However, these model-based approaches cannot be used when working with black-box machine learning models or ad hoc clustering algorithms.

In contrast, our approach to estimation is more practical than pairwise sampling and applies to any black-box disambiguation algorithms such as those used at

PatentsView.

3.1.1.3 Our Approach

Our approach to estimating performance metrics is based on the use of benchmark datasets that already exist or that can be collected in a cost-effective way. These datasets contain entity clusters corresponding to either: (a) sampling records and recovering all associated instances, (b) directly sampling clusters, or (c) sampling blocks. For each of these sampling processes, we propose estimators that correct for the issues discussed in section 3.1.1, are nearly unbiased, and are easy to use in practice.

Our approach has the following advantages:

1. It can leverage existing benchmark datasets as well as new datasets collected specifically for performance evaluation.
2. It can easily be generalized to estimate other clustering metrics, such as cluster precision, cluster recall, cluster homogeneity and completeness, and other generalized merge distances.
3. For evaluation, the review of entity clusters is much more efficient than the review of record pairs. We can achieve high accuracy with small samples without relying on sophisticated sampling schemes.

Furthermore, our approach is novel. To our knowledge, we are the first to propose unbiased performance estimators based on cluster and block samples. Past work either ignored biases when computing precision and recall from benchmark datasets (Frisoli and Nugent, 2018; Monath et al., 2021; Han et al., 2019), did not provide estimates for precision or recall (McVeigh et al., 2019), or provided solutions tailored to very specific record linkage models (Belin and Rubin, 1995). We provide the first

general solution to entity resolution evaluation that does not rely on sampling record pairs and that applies to any disambiguation algorithm.

In short, the proposed approach is simple, principled, and practical. It is simple to use, it is statistically principled in its account of sampling processes and uncertainty, and it is practical in the way that it can provide cost-effective estimates for any disambiguation algorithm.

3.1.2 Structure of the Paper

The rest of the paper is organized as follows. In section 3.2, we describe benchmark datasets, our hand-disambiguation methodology for evaluation, the proposed estimators, and our simulation study that we use to validate the performance of our estimators. Section 3.3 then presents our performance estimates and results from the simulation study. Section 3.4 summarizes the paper and explores future research directions.

3.2 Data and Methodology

In this section, we introduce the benchmark datasets used at PatentsView, our hand-disambiguation methodology, our proposed performance metric estimators, and the simulation framework that we use to compare estimators. Note that we focus on inventor disambiguation throughout, rather than on the related problems of assignee and location disambiguation.

3.2.1 Benchmark Datasets for Inventor Disambiguation

We consider the following benchmark datasets for inventor disambiguation.

3.2.1.0.1 Israeli Inventors Benchmark

Trajtenberg and Shiff (2008) disambiguated the U.S. patents of Israeli inventors

that were granted between 1963 and 1999. A total of 6,023 Israeli inventors were identified for this time period with 15,310 associated patents.

3.2.1.0.2 Li et al. (2014)’s Inventors Benchmark

Based on an original dataset from Gu et al. (2008), Li et al. (2014) disambiguated the patent history (between 1975 and 2010) of 95 U.S. inventors.

3.2.2 Hand-Disambiguation Methodology

In addition to considering the above benchmark datasets, we have carried out hand-disambiguation of inventor mentions. This was motivated by the evaluation of the current PatentsView inventor disambiguation using the estimators proposed in section 3.2.3.

In total, 100 inventors were sampled with probability proportional to their number of granted patents. This was done by sampling inventor mentions uniformly at random and recovering all patents for a given inventor. These inventor mentions were from U.S. patents granted between 1976 and December 31, 2021.

Two AIR staff were tasked with recovering inventors’ patents given sampled inventor mentions. First, given a sampled inventor, the associated predicted cluster was reviewed and any wrongly assigned patents were removed. Next, PatentsView’s search tools were used to find additional mentions of similarly named inventors. These inventor mentions were reviewed and added to the predicted cluster, if appropriate. The two AIR staff had an initial training session, followed by a test run on 10 inventors, before carrying out the rest of the data collection. They worked independently, which resulted in two datasets being obtained for the same inventor mentions. In section 3.3, these are referred to as the **Staff 1** and **Staff 2** datasets.

Note that our data collection methodology is biased toward PatentsView’s current disambiguation. Indeed, we did not expect the staff to have found all errors or all

missing inventor mentions from the predicted clusters. The staff used their best judgment, supported by a thorough search, to resolve inventor mentions. In cases where no errors were found, the current disambiguation was assumed to be correct. Performance estimates based on this data might therefore be slightly optimistic, which should be acknowledged when reporting performance estimates to PatentsView.org users. Otherwise, for the purpose of improving the current disambiguation algorithm, this data is still appropriate to use. It represents the most visible errors in the current disambiguation rather than the totality of them.

3.2.3 Proposed Performance Estimators

Throughout the rest of the paper, we focus on pairwise precision and pairwise recall (defined below in (3.1)) as our performance evaluation metrics.

3.2.3.1 Representation Lemmas

First, we define pairwise precision and recall in terms of the number of links between records. Let $\mathcal{D} = \{1, 2, 3, \dots, N\}$ index a set of records let \mathcal{C} be the partition of \mathcal{D} representing ground truth clustering, and let $\hat{\mathcal{C}}$ be a set of predicted clusters. Now let \mathcal{T} be the set of record pairs that appear in the same cluster in \mathcal{C} (matching pairs), and let \mathcal{P} be the set of record pairs that appear in the same predicted cluster in $\hat{\mathcal{C}}$ (predicted links). Pairwise precision (P) and pairwise recall (R) are then defined as

$$P = \frac{|\mathcal{T} \cap \mathcal{P}|}{|\mathcal{P}|}, \quad R = \frac{|\mathcal{T} \cap \mathcal{P}|}{|\mathcal{T}|}. \quad (3.1)$$

Note that $P = R|\mathcal{T}|/|\mathcal{P}|$. As such, precision and recall are equal if and only if the right number of matching pairs is predicted under $\hat{\mathcal{C}}$.

We now provide three alternative representations of precision and recall that correspond to the processes of sampling records, sampling true clusters, and sampling

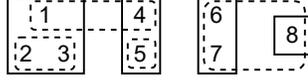


FIGURE 3.2: Example of ground truth clustering \mathcal{C} (represented by boxes with a full border) and predicted clustering $\hat{\mathcal{C}}$ (rounded boxes with a dotted border) of elements $\mathcal{D} = \{1, 2, \dots, 8\}$. Here $\mathcal{T} = \{(1, 2), (2, 3), (1, 3), (4, 5), (6, 7)\}$ and $\mathcal{P} = \{(1, 4), (2, 3), (6, 7), (7, 8), (6, 8)\}$. As such, $P = R = 2/5$ in this example.

blocks. These representations will be used to obtain precision and recall estimators under these sampling processes.

3.2.3.1.1 Record Sampling Representation

For a given $i \in \mathcal{D}$, let $c(i) \in \mathcal{C}$ be the ground truth cluster associated with i in \mathcal{C} .

For a given $c \in \mathcal{C}$, we define

$$f(c, \hat{\mathcal{C}}) = \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|c \cap \hat{c}|}{2}, \quad g(c, \hat{\mathcal{C}}) = \frac{f(c, \hat{\mathcal{C}})}{\sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|\hat{c}|}{2}}. \quad (3.2)$$

Lemma 1. If i is distributed over \mathcal{D} with probabilities $p_i > 0$, then

$$P = \mathbb{E} \left[\frac{g(c(i), \hat{\mathcal{C}})}{p_i |c(i)|} \right], \quad R = 2 \frac{\mathbb{E} \left[\frac{f(c(i), \hat{\mathcal{C}})}{|c(i)| p_i} \right]}{\mathbb{E} [(|c(i)| - 1)/p_i]}. \quad (3.3)$$

Proof. By breaking down \mathcal{P} and \mathcal{T} over predicted clusters, we find

$$P = \sum_{c \in \mathcal{C}} \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|c \cap \hat{c}|}{2} / \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|\hat{c}|}{2}. \quad (3.4)$$

Now writing $\sum_{c \in \mathcal{C}} \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|c \cap \hat{c}|}{2} = \sum_{i=1}^N \frac{1}{|c(i)|} \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|c(i) \cap \hat{c}|}{2}$ and substituting $g(c(i), \hat{\mathcal{C}})$, we obtain

$$P = \sum_{i=1}^N p_i \frac{g(c(i), \hat{\mathcal{C}})}{p_i |c(i)|} = \mathbb{E} \left[\frac{g(c(i), \hat{\mathcal{C}})}{p_i |c(i)|} \right]. \quad (3.5)$$

For recall, write

$$R = \sum_{c \in \mathcal{C}} \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|c \cap \hat{c}|}{2} / \sum_{c \in \mathcal{C}} \binom{|c|}{2}. \quad (3.6)$$

Through a similar argument as above, we may express the numerator as

$$\sum_{c \in \mathcal{C}} \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|c \cap \hat{c}|}{2} = \mathbb{E} \left[f(c(i), \hat{\mathcal{C}}) / (|c(i)| p_i) \right]. \quad (3.7)$$

For the denominator, we have

$$\sum_{c \in \mathcal{C}} \binom{|c|}{2} = \sum_{i=1}^N \frac{1}{|c(i)|} \binom{|c(i)|}{2} = \mathbb{E}[(|c(i)| - 1) / (2p_i)]. \quad (3.8)$$

Combining (3.7) and (3.8) yields the result. \square

3.2.3.1.2 Cluster Sampling Representation

In the cluster sampling case, sampling probabilities are typically known only up to a normalizing factor. This is because the total number of true clusters and other aspects of the ground truth cluster distribution are unknown in practice. As such, we provide expressions for precision and recall that only require knowing the sampling probabilities up to a normalizing factor. This allows the consideration of sampling uniformly at random and sampling clusters with probability proportional to their size.

Lemma 2. If c is distributed over \mathcal{C} with probabilities proportional to $p_c > 0$, then

$$P = \frac{N \mathbb{E} \left[g(c, \hat{\mathcal{C}}) / p_c \right]}{\mathbb{E} \left[|c| / p_c \right]}, \quad R = \frac{\mathbb{E} \left[f(c, \hat{\mathcal{C}}) / p_c \right]}{\mathbb{E} \left[\binom{|c|}{2} / p_c \right]}. \quad (3.9)$$

Proof. Let $\pi > 0$ be such that $\sum_{c \in \mathcal{C}} \pi p_c = 1$. Now write

$$P = \sum_{c \in \mathcal{C}} g(c, \hat{\mathcal{C}}) = |\mathcal{C}| \sum_{c \in \mathcal{C}} \pi p_c g(c, \hat{\mathcal{C}}) / (\pi p_c |\mathcal{C}|) = |\mathcal{C}| \mathbb{E} \left[g(c, \hat{\mathcal{C}}) / (\pi p_c |\mathcal{C}|) \right] \quad (3.10)$$

and

$$|\mathcal{C}| = \frac{N}{\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} |c|} = \frac{N}{\mathbb{E}[|c|/(\pi p_c |\mathcal{C}|)]}. \quad (3.11)$$

Simplifying $\pi|\mathcal{C}|$ from the numerator and denominator then yields the expression for precision.

The expression for recall follows in a straightforward way from (3.6). \square

Note that, with clusters sampled with probability proportional to their size, the expression for precision simplifies to $P = N \mathbb{E} \left[g(c, \hat{\mathcal{C}}) / |c| \right]$.

Remark 5. Lemma 1 and lemma 2 can be generalized to apply to any performance metric that can be expressed as a sum $\sum_{c \in \mathcal{C}} h(c, \hat{\mathcal{C}})$, for some function h , or as a function of such sums. For instance, the use of the so-called cluster precision and cluster recall (Barnes, 2015), or of cluster homogeneity and completeness (Barnes, 2015), can be more appropriate in the presence of large clusters. We leave these generalizations as extensions of our work.

3.2.3.1.3 Disjoint Block Sampling Representation

Let \mathcal{B} be a partition of \mathcal{D} such that for every $c \in \mathcal{C}$, there exists $b \in \mathcal{B}$ with $c \subset b$. For a given $b \in \mathcal{B}$, let \mathcal{T}_b be the set of ground truth links contained within b , let \mathcal{P}_b be the set of predicted links contained within b , and let \mathcal{P}_b^- be the set of predicted links with a single record in b (i.e., \mathcal{P}_b^- is the set of outgoing links from b).

Lemma 3. If b is distributed over \mathcal{B} with probabilities proportional to $p_b > 0$, then

$$P = \frac{\mathbb{E}[|\mathcal{T}_b \cap \mathcal{P}_b|/p_b]}{\mathbb{E}[(|\mathcal{P}_b| + \frac{1}{2}|\mathcal{P}_b^-|)/p_b]}, \quad R = \frac{\mathbb{E}[|\mathcal{T}_b \cap \mathcal{P}_b|/p_b]}{\mathbb{E}[|\mathcal{T}_b|/p_b]}. \quad (3.12)$$

Proof. Since the blocking procedure is assumed to have no error (for every $c \in \mathcal{C}$, there exists $b \in \mathcal{B}$ with $c \subset b$), we can break down \mathcal{T} as the disjoint union of the

\mathcal{T}_b 's over $b \in \mathcal{B}$. It follows that $|T| = \sum_{b \in \mathcal{B}} |\mathcal{T}_b|$ and $|\mathcal{T} \cap \mathcal{P}| = \sum_{b \in \mathcal{B}} |\mathcal{T}_b \cap \mathcal{P}_b|$. The expression for recall in (3.12) follows directly. For precision, we can express $|\mathcal{P}|$ as the number of links within blocks plus the number of links across blocks. Since the number of links across blocks is counted twice when each block is considered, we obtain $|\mathcal{P}| = \sum_{b \in \mathcal{B}} (|\mathcal{P}_b| + \frac{1}{2}|\mathcal{P}_b^-|)$. \square

Remark 6. Lemmas 1 – 3 are formulated in terms of sampled ground truth clusters and sampled blocks that do not contain errors. However, given the duality between precision and recall (interchanging the roles between \mathcal{C} and $\hat{\mathcal{C}}$ interchanges precision and recall), the results also apply to sampling predicted clusters.

3.2.3.2 Proposed Estimators

All of the expressions for precision and recall in lemmas 1 – 3 are either population means or ratios of population means. As such, they can be estimated using sample means and ratios of sample means. For readability, we present here a generic formula for an approximately unbiased estimator of the ratio of means and then specify the needed quantities for each representation below. The estimator applies a first order bias correction to the ratio of sample means, based on a Taylor approximation. Approximate confidence intervals can be computed based on the variance estimator of the ratio of sample means by Taylor approximation, assuming the corrected estimator has a small bias (Särndal et al., 2003; Fuller, 2011). The generic formula for estimating the ratio of T -sized “population” (of records/clusters/blocks) means of the form

$$E = \frac{1/T \sum_{i=1}^T B_i}{1/T \sum_{i=1}^T A_i}, \quad (3.13)$$

assuming we have sampled n elements (records/clusters/blocks), is

$$\hat{E} = \frac{\bar{B}_n}{\bar{A}_n} \left\{ 1 + \frac{\theta_{n,T}}{n(n-1)} \sum_{s=1}^n \frac{A_s}{\bar{A}_n} \left(\frac{B_s}{\bar{B}_n} - \frac{A_s}{\bar{A}_n} \right) \right\}, \quad \bar{A}_n = \frac{1}{n} \sum_{s=1}^n A_s, \quad \bar{B}_n = \frac{1}{n} \sum_{s=1}^n B_s \quad (3.14)$$

We note that an additional symbol $\theta_{n,T}$ is introduced for a possible finite population correction when relatively large number of elements are sampled without replacement (see below). Classical adjustment is set to $\theta_{n,T} = (1 - \frac{n-1}{T-1})$ (Cochran, 1977). For practical purposes when T is large, $\theta_{n,T} = 1$ will suffice. In fact, knowledge of T is not needed at all as long as it is large enough relative to n , which is useful because the total number of true clusters/blocks is not known in advance. Confidence intervals can be computed based on the variance estimate of the above, which is

$$\hat{V}(\hat{E}) = \left(\frac{\bar{B}_n}{\bar{A}_n}\right)^2 \frac{\theta_{n,T}}{n(n-1)} \sum_{s=1}^n \left(\frac{A_s}{\bar{A}_n} - \frac{B_s}{\bar{B}_n}\right)^2. \quad (3.15)$$

The specific values for A_s , B_s , and T in each of our representation are described in appendix 3.5.2.

Remark 7. In entity resolution applications, we can typically assume that elements have been sampled with replacement (or closely so). Indeed, with a small proportion of sampled elements, nonreplacement samples are approximately equivalent to samples with replacement. However, if dealing with relatively large nonreplacement samples, then the sampling probabilities used in the definition of the estimators should be adjusted to reflect the size-dependent effect of nonreplacement (Horvitz and Thompson, 1952).

3.2.4 Simulation Study

In order to assess the performance of the proposed estimators, we carried out a simulation study based on PatentsView’s inventor disambiguation. Specifically, in the context of the simulation, we considered PatentsView current inventor disambiguation as the ground truth clustering. A simulated set of predicted clusters was obtained by introducing errors (misattribution of inventor mentions) into the current disambiguation. We then estimated the precision and recall of this predicted clustering using our

estimators based on random cluster samples. The process of sampling clusters and estimating precision/recall was repeated 100 times in order to provide the distribution of the estimators and metrics such as bias and root mean squared error (rmse).

To introduce errors, we picked records at random and changed their cluster assignment to that of other records picked at random. This is a simple process that ensures that larger clusters are more likely to contain errors. In our simulation, we considered rates of 5%, 10%, and 25% for the proportion of records that are sampled for cluster misassignment. Although the larger error rates are more realistic, the 5% misattribution rate helps showcase the properties of our estimators when only a small proportion of the sampled clusters is associated with errors.

For the sampling process, we considered sampling records uniformly at random and recovering their associated clusters. This is the same as sampling clusters with probability proportional to their size. In the record/cluster sampling cases, we looked at the effect of sampling 100, 200 and 400 records/clusters.

Finally, we compared the following three precision and recall estimators:

P_naive, R_naive This is the “naive” precision (respectively recall) estimator obtained by computing precision (respectively recall) when only looking at records that appear in the sampled clusters.

P_record, R_record These are the precision and recall estimators corresponding to uniformly sampling records in lemma 1 ($p_i \propto 1$) with the bias adjustment given in (3.14). Note that these are the same as the estimators obtained from lemma 2 when sampling clusters with probability proportional to their size ($p_c \propto |c|$).

P_cluster_block This is the precision estimator obtained by considering each sampled cluster as its own block in lemma 3, where clusters have been sampled with probability proportional to their size ($p_b \propto |b|$) and with the bias adjustment

given in (3.14). Note that in the case of recall with cluster blocks, the estimator corresponding to lemma 3 is the same as the one corresponding to lemma 2.

3.3 Results

3.3.1 Results From the Simulation Study

Figure 3.3 shows the distribution of the three precision estimators used in the simulation study (see section 3.2.4) compared to ground truth precision. The block sampling estimator `P_cluster_block` is highly accurate, while `P_record` is more variable and `P_naive` is entirely uninformative. Note that `P_record` can take values greater than 1 (not shown in this figure), so that truncating it to be less than 1 introduces a bias in some cases.

`P_cluster_block` performs better than `P_record` because `P_record` has been derived in a generic way that applies to any performance metric that can be expressed in cluster form similar to (3.9). On the other hand, `P_cluster_block` relies on specific properties of precision. Among other things, this ensures that `P_cluster_block` is constrained to be between 0 and 1. In practice, `P_cluster_block` should be preferred as a pairwise precision estimator. The bias and rmse of the precision estimators are reported in table 3.1.

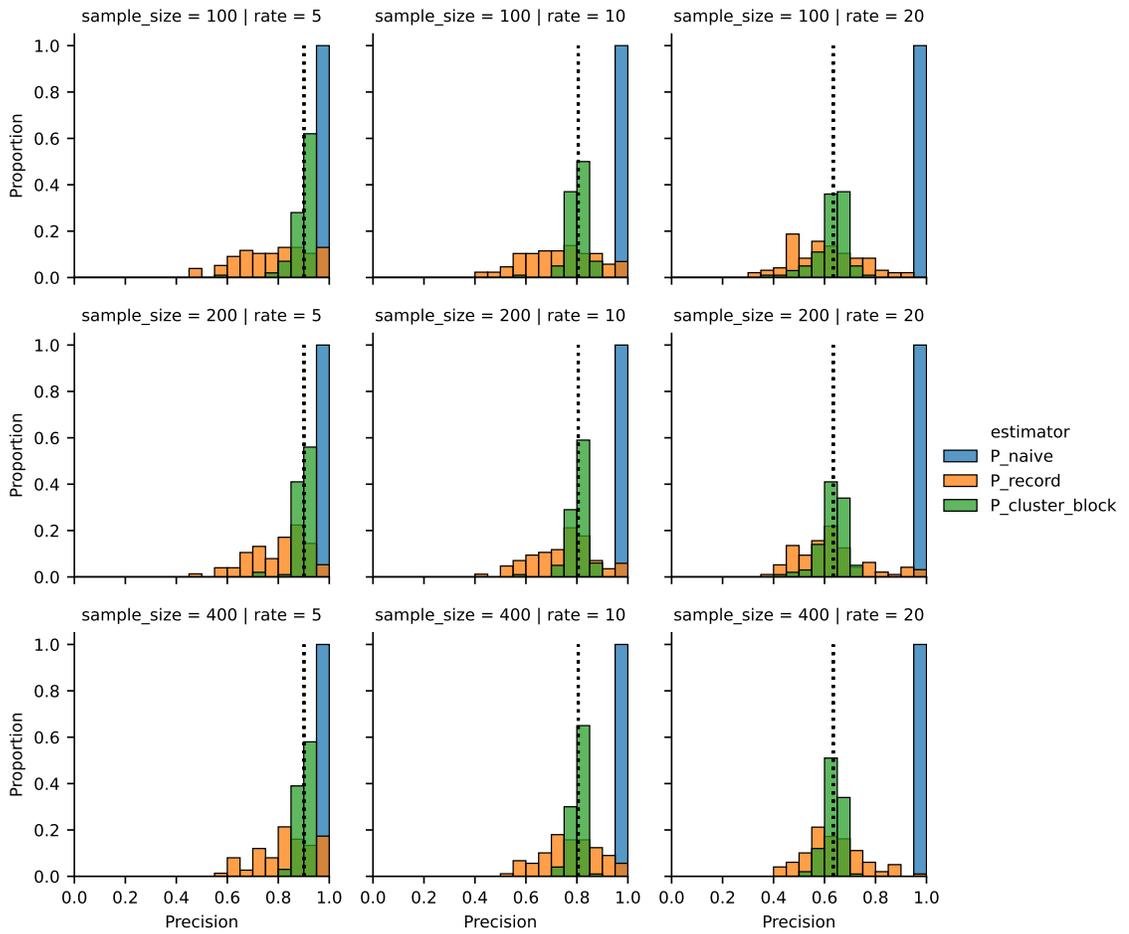


FIGURE 3.3: Distribution of precision estimates for various sample sizes and misattribution rates, as described in section 3.2.4. Ground truth precision is marked by a dotted vertical line. The `rate` variable represents the percentage misattribution rate. The estimator `P_cluster_block` is highly accurate, while `P_naive` is almost always close to 1.0, having little to do with the true precision.

Table 3.1: Bias and root mean squared error (rmse) of precision estimators for the simulation study described in section 3.2.4. The `rate` variable represents the percentage misattribution rate.

		rate			5			10			20		
		sample size			100	200	400	100	200	400	100	200	400
bias	P_cluster_block	-0.004	-0.002	-0.001	-0.002	-0.001	0.001	-0.004	-0.001	-0.000			
	P_naive	0.099	0.099	0.099	0.195	0.195	0.195	0.366	0.366	0.365			
	P_record	-0.002	0.013	0.009	-0.001	0.011	0.008	-0.001	0.009	0.006			
rmse	P_cluster_block	0.045	0.033	0.021	0.041	0.038	0.026	0.063	0.052	0.034			
	P_naive	0.099	0.099	0.099	0.195	0.195	0.195	0.366	0.366	0.365			
	P_record	0.294	0.232	0.169	0.260	0.205	0.150	0.207	0.162	0.117			

Regarding recall, figure 3.4 shows the distribution of the two estimators used in the simulation study. The naive recall estimator performs well in this case. However, the recall estimator accounting for the sampling mechanism is more accurate (**R_record**, which is equal to **R_cluster_block**). The bias and rmse of the recall estimators are reported in table 3.2, where the overall improved performance of **R_record** can be observed.

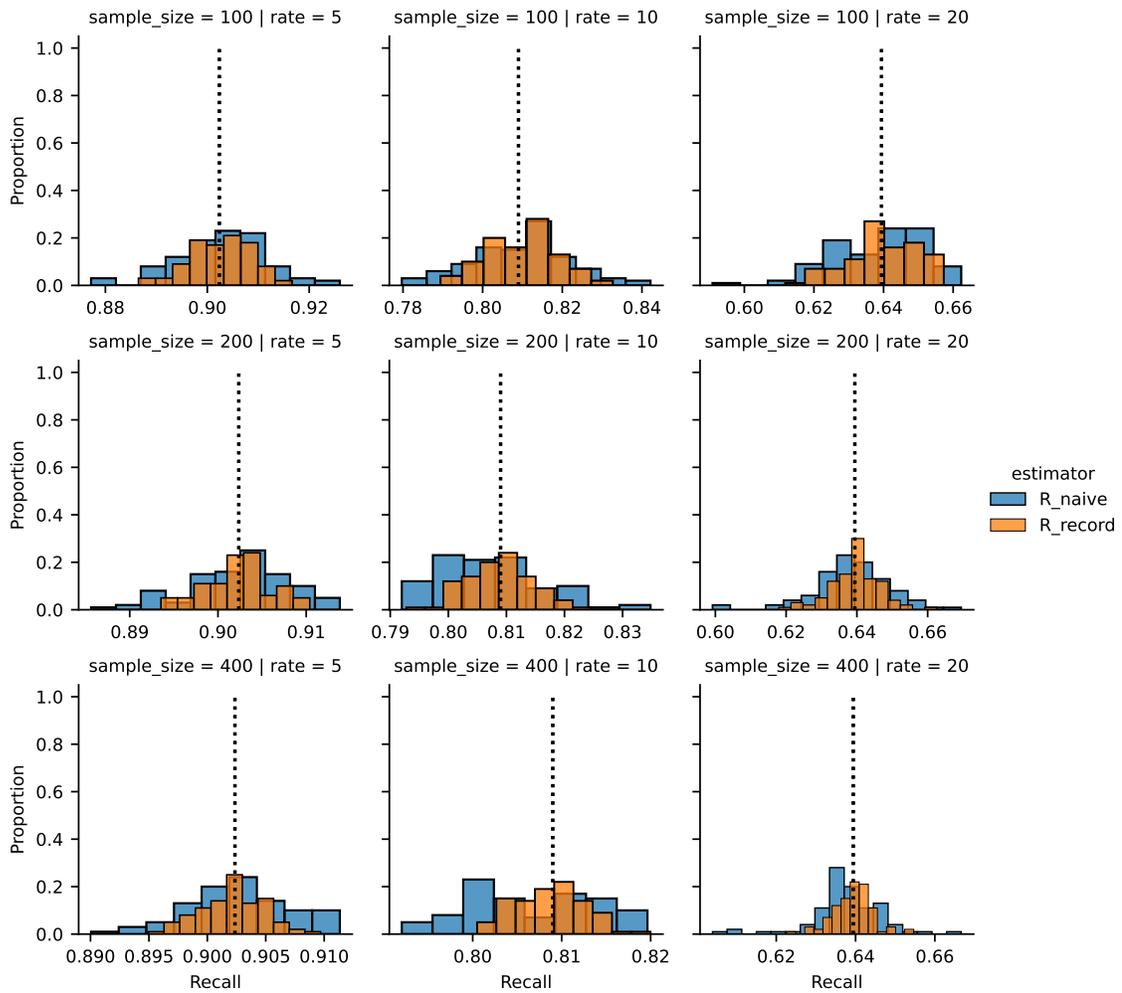


FIGURE 3.4: Distribution of recall estimates for various sample sizes and misattribution rates, as described in section 3.2.4. Ground truth recall is marked by a dotted vertical line. The `rate` variable represents the percentage misattribution rate.

Table 3.2: Bias and root mean squared error of recall estimators for the simulation study described in section 3.2.4. The `rate` variable represents the percentage misattribution rate.

		5			10			20		
	rate									
	sample size	100	200	400	100	200	400	100	200	400
bias	R_record	0.0002	-0.0001	-0.0002	0.0011	-0.0002	-0.0004	0.0005	0.0001	-0.0002
	R_naive	0.0007	0.0001	0.0004	0.0009	-0.0015	-0.0024	-0.0008	-0.0010	-0.0017
rmse	R_record	0.0060	0.0037	0.0028	0.0089	0.0056	0.0039	0.0099	0.0071	0.0050
	R_naive	0.0089	0.0056	0.0043	0.0123	0.0089	0.0075	0.0132	0.0109	0.0091

3.3.2 Evaluation of PatentsView’s Disambiguation

Table 3.3 shows estimated pairwise precision and recall from benchmark datasets and from our two hand-curated datasets. Note that each estimate is associated with a given population of inventor mentions which is a subset of granted U.S. patents since 1976. We focused on U.S. patents granted since 1976 as this is the main data product of PatentsView.org.

The choice of estimators for the results presented in table 3.3 is as follows. Since our two hand-curated datasets (see section 3.2.2) were obtained by sampling inventor clusters with probabilities proportional to cluster sizes, we used the `P_cluster_block` and `R_record` estimators with corresponding probability weights. For the Israeli benchmark dataset, we assumed that a single block of inventor clusters was sampled and used the corresponding estimators defined through (3.12) with a single block sample. Note that no variance estimates can be given for single samples. For Li et al. (2014)’s inventors benchmark, given that the inventor clusters were originally obtained from a set of inventor curriculum vitae, we assumed that the clusters were sampled uniformly at random. As such, we used cluster block estimators with constant probability weights.

Table 3.3: Estimated pairwise precision and recall (with estimated standard deviation) from our four benchmark datasets.

dataset	est. precision ($\hat{\sigma}$)	est. recall ($\hat{\sigma}$)	scope
Staff 1	88% (3.4%)	95% (1.1%)	1976 – Dec. 31, 2022 (U.S. granted)
Staff 2	87% (3.6%)	96% (1.0%)	1976 – Dec. 31, 2022 (U.S. granted)
Israeli Benchmark	79% (NA)	94% (NA)	1976 – 1999 (U.S. granted)
Li et al. (2014)’s Benchmark	91% (2.7%)	91% (5.0%)	1976 – 2010 (U.S. granted)

Overall, our performance estimates paint the first realistic picture of PatentsView’s disambiguation accuracy in practice. Precision is not nearly 100%, as would be assumed from naively computing precision on benchmark datasets. Rather, there is

significant room for improvement. Our hand-curated datasets and data collection methodology provide the necessary basis to investigate errors and plan for improvements to the disambiguation algorithm.

3.4 Discussion

Motivated by PatentsView’s disambiguation, this paper introduced a novel evaluation methodology for entity resolution algorithms. The methodology relies on benchmark datasets containing ground truth clusters and estimators that account for biases inherent to these datasets. For PatentsView, this provided the first representative estimates of its disambiguation performance. Furthermore, all data and code used in this paper, as well as other tools developed to facilitate evaluation at PatentsView, are freely available at <https://github.com/PatentsView/PatentsView-Evaluation/>.

There are two main products resulting from this work. The first is the appropriate understanding of the quality of the data provided to PatentsView’s users. Our performance estimates indicate that, despite an overall accurate disambiguation, there is significant room for improvement. Notably, the current disambiguation over-estimates the number of matching inventor mention pairs. The second product is the set of tools needed for methodological research and model comparison. Given our evaluation methodology, we can now reliably compare algorithms and decide with confidence on changes that will affect users.

One important topic for future work is the quantification of uncertainty associated with errors in the hand-disambiguation process. This is a challenging problem given the lack of validation information available. Surveying inventors to validate the hand-disambiguation process would be one way to explore this issue. Sensitivity analyses involving potential errors in the hand-disambiguation process could also be informative. We refer the reader to Bailey et al. (2017) for a state-of-the-art

hand-labeling study that evaluated human review accuracy.

Another important topic for future work is the development of estimators for additional performance metrics. As we have seen in the simulation study, the generically derived estimators (e.g., from lemma 2) do not perform as well as estimators derived using specific properties of pairwise precision and recall (lemma 3). As such, care should be taken to obtain efficient estimators for every metric of interest.

Finally, we note that the performance of estimators can degrade when dealing with heavy-tailed cluster size distributions. The bias of ratio estimators can be high in this case, especially when using small samples and when sampling clusters uniformly at random rather than with probability proportional to size. Model-based estimators that exploit known properties of the cluster size distribution could be developed to improve estimation accuracy in such cases.

Data and Code

All data and code used for this paper are available as part of the PatentsView-Evaluation Python package at <https://github.com/PatentsView/PatentsView-Evaluation/>.

Author Contributions

Olivier Binette led the evaluation project and wrote most of this manuscript. Sokhna A York and Emma Hickerson carried out the data collection by manually reviewing inventor clusters. Youngsoo Baek provided bias adjustment and uncertainty quantification for ratio estimators. Sarvo Madhavan was a technical advisor and contributed to code. Christina Jones was an advisor and project manager. All authors provided input on the manuscript.

3.5 Appendix

3.5.1 Bias of Precision Computed on Benchmark Datasets

This section provides more information on example 1.

For this example, we considered the RLdata10000 dataset from Sariyar and Borg (2022). This is a synthetic dataset containing 10,000 records with first name, last name, and date of birth attributes. There is noise in these attributes and a 10% duplication rate. Ground truth identity is known for all records.

The disambiguation algorithm we consider matches records if any of the following conditions are met:

- records agree on first name, last name, and birth year,
- records agree on first name, birth day, and birth year, or
- records agree on last name, birth day, and birth year.

Note that this is not at all a good disambiguation algorithm. It has 52% precision and 83% recall. However, it allows us to showcase the issue with nonadjusted precision computed on cluster samples.

In our experiment, we have repeated 5,000 times the following three-steps process 5,000 times:

1. First, 200 records were sampled and the ground truth clusters associated with them were recovered. This step provided a "benchmark" dataset that was used for evaluation.
2. Second, a trivial precision estimate was obtained by computing precision over the benchmark dataset. That is, predicted cluster assignments were restricted to records that appear in the benchmark data and precision was compared for

these records. More often than not, the result was an observation of 100% precision.

3. Third, we computed our proposed precision estimator which corresponds to lemma 1 and the estimator \hat{P}_{block} defined in (3.21), with blocks corresponding to clusters and with sampling probabilities $p_b \propto |b|$.

The distributions of the two precision estimates over the 5,000 repetitions are shown in figure 3.1. Our proposed estimator is accurate and nearly unbiased, whereas the trivial precision estimates have almost nothing to do with actual algorithmic performance.

3.5.2 Precision and Recall Estimator Formulas

This section describes specific values for A_s , B_s , and T in (3.14) in order to obtain nearly unbiased precision and recall estimators based on each of the representations in section 3.2.3. We use the symbols \hat{P} and \hat{R} , indexed by either "rec", "clust", or "block", in order to refer to precision and recall estimators corresponding to record, cluster, and block sampling representations, respectively.

3.5.2.0.1 Record Sampling Estimators

\hat{P}_{rec} Estimating P is a special case that does not require ratio-of-means estimation.

We propose to use a simple unbiased estimator:

$$\hat{P}_{\text{rec}} = \frac{1}{n} \sum_{s=1}^n \frac{g(c(i_s), \hat{\mathcal{C}})}{|c(i_s)| p_{i_s}}. \quad (3.16)$$

The unbiased estimator of the variance of \hat{P}_{rec} is also available:

$$\hat{V}(\hat{P}_{\text{rec}}) = \frac{\theta_{n,N}}{n(n-1)} \sum_{s=1}^n \left(\frac{g(c(i_s), \hat{\mathcal{C}})^2}{|c(i_s)|^2 p_{i_s}^2} - \hat{P}_{\text{rec}}^2 \right) \quad (3.17)$$

\widehat{R}_{rec} Going forward, we refer to formulae (3.14) and (3.15). \widehat{R}_{rec} and its variance estimate $\widehat{V}(\widehat{R}_{\text{rec}})$ are given by (3.14) and (3.15), where we substitute in

$$T = N, A_s = \frac{(|c(i_s)| - 1)}{p_{i_s}}, B_s = 2 \frac{f(c(i_s), \widehat{\mathcal{C}})}{|c(i_s)| p_{i_s}}. \quad (3.18)$$

3.5.2.0.2 Cluster Sampling Estimators

$\widehat{P}_{\text{clust}}$ The estimator and its variance estimate $\widehat{V}(\widehat{P}_{\text{clust}})$ are given by (3.14) and (3.15), where we substitute in

$$T = |\mathcal{C}|, A_s = \frac{|c_s|}{p_{c_s}}, B_s = N \frac{g(c_s, \widehat{\mathcal{C}})}{p_{c_s}}. \quad (3.19)$$

$\widehat{R}_{\text{clust}}$ The estimator and its variance estimate $\widehat{V}(\widehat{R}_{\text{clust}})$ are given by (3.14) and (3.15), where we substitute in

$$T = |\mathcal{C}|, A_s = \frac{\binom{|c_s|}{2}}{p_{c_s}}, B_s = \frac{f(c_s, \widehat{\mathcal{C}})}{p_{c_s}}. \quad (3.20)$$

3.5.2.0.3 Disjoint Block Sampling Estimators

$\widehat{P}_{\text{block}}$ The estimator and its variance estimate $\widehat{V}(\widehat{P}_{\text{block}})$ are given by (3.14) and (3.15), where we substitute in

$$T = |\mathcal{B}|, A_s = \frac{|\mathcal{P}_{b_s}| + \frac{1}{2}|\mathcal{P}_{b_s}^-|}{p_{b_s}}, B_s = \frac{|\mathcal{T}_{b_s} \cap \mathcal{P}_{b_s}|}{p_{b_s}}. \quad (3.21)$$

$\widehat{R}_{\text{block}}$ The estimator and its variance estimate $\widehat{V}(\widehat{R}_{\text{block}})$ are given by (3.14) and (3.15), where we substitute in

$$T = |\mathcal{B}|, A_s = \frac{|\mathcal{T}_{b_s}|}{p_{b_s}}, B_s = \frac{|\mathcal{T}_{b_s} \cap \mathcal{P}_{b_s}|}{p_{b_s}}. \quad (3.22)$$

4. How to Evaluate Entity Resolution Systems: An Entity-Centric Framework with Application to Inventor Name Disambiguation

Entity resolution (record linkage, microclustering) systems are notoriously difficult to evaluate. Looking for a needle in a haystack, traditional evaluation methods use sophisticated, application-specific sampling schemes to find matching pairs of records among an immense number of non-matches. We propose an alternative that facilitates the creation of representative, reusable benchmark data sets without necessitating complex sampling schemes. These benchmark data sets can then be used for model training and a variety of evaluation tasks. Specifically, we propose an entity-centric data labeling methodology that integrates with a unified framework for monitoring summary statistics, estimating key performance metrics such as cluster and pairwise precision and recall, and analyzing root causes for errors. We validate the framework in an application to inventor name disambiguation and through simulation studies. Software: <https://github.com/OlivierBinette/er-evaluation/>

4.1 Introduction

Entity resolution is the process of identifying and linking database records referring to the same entity, such as a person or organization (Christen and Christen, 2012; Christophides et al., 2021; Papadakis et al., 2021; Binette and Steorts, 2022a). In the

absence of a reliable unique identifier, this is a large-scale clustering task: records need to be grouped into clusters, each representing a unique entity. In many applications, these clusters are numerous yet small. For example, entity resolution is used for the identification of unique inventors listed on U.S. Patents and Trademarks Office (USPTO) patents. This must account for commonalities in inventor names and the frequent occurrence of errors and variations in recorded names (Li et al., 2014). The task involves resolving millions of unique inventors, many of whom have authored only a handful of patents, while others may have contributed to hundreds.

Entity resolution systems typically employ machine learning and artificial intelligence models to tackle this challenge. The models use contextual information to predict whether or not two records refer to the same entity. For example, we can approximately resolve unique inventors by using patent topic, co-authors, employer, and location, in addition to first and last names (Monath et al., 2021). A large body of research has considered this problem, as this is an error-prone process that is further complicated by the scale of the clustering task (Li et al., 2014; Huberty et al., 2014; Pezzoni et al., 2014; Balsmeier et al., 2015; Ventura et al., 2015; Kim et al., 2016a; Yang et al., 2017; Doherr, 2017; Han et al., 2019; Yin et al., 2020; Monath et al., 2021; Doherr, 2021b).

In this article, we address evaluating the accuracy of entity resolution systems. Traditional evaluation methods often rely on manually reviewing pairs of records to validate linkage predictions. However, finding matching pairs, especially those missed by the entity resolution system, is much like looking for a needle in a haystack: in a database of n records, there are $\mathcal{O}(n^2)$ non-matches and only $\mathcal{O}(n)$ matches. Even if these matches can be found using a smart sampling scheme built around a specific entity resolution system, the resulting data are not necessarily well-suited for evaluating or training other models, or for estimating other metrics besides pairwise precision and recall.

Instead of reviewing pairs of records, our evaluation approach utilizes a sample of fully-resolved entities, i.e., ground truth, or known clusters. In this approach, all pairs within a resolved cluster are known to match, and any pair that intersects a resolved cluster but is not contained within it is known to be a non-match. For example, a resolved cluster of 10 records in a database of 1,000,010 records includes $\binom{10}{2} = 45$ matching pairs and excludes 10 million non-matching pairs. We show that sampling ground truth clusters, and using the resulting matches and non-matches, facilitates the estimation of performance metrics without having to rely on sophisticated pairwise sampling schemes that find sufficient numbers of matching pairs among the massive number of nonmatching pairs (e.g., as in Marchant and Rubinstein, 2017).

In order to use fully-resolved entities as the starting point of evaluation, we propose an entity-centric evaluation framework with the following components.

- 1. Cluster-Wise Error Metrics:** To identify errors made by an entity resolution system, we compare predicted clusters against a sample of known, fully-resolved clusters through error metrics defined at the record and cluster levels (section 4.3.3.1).
- 2. Global Performance Metric Estimates:** To obtain estimates of global performance metrics such as pairwise and b-cubed precision and recall (Michelson and Macskassy, 2009; Menestrina et al., 2010; Barnes, 2015), we express them as weighted aggregates of cluster-wise error metrics. This helps obtain estimates that are representative of the system’s performance on the entire data set, not just the benchmark (section 4.3.4).
- 3. Error Analysis:** To analyze the root causes of errors, we relate errors to entity features extracted from resolved clusters of records. (section 4.3.3.2).

Furthermore, to support this process, we introduce:

4. Data Labeling Through Cluster Sampling: A methodology for creating a benchmark set of fully-resolved entities through manual data labeling (section 4.3.2).

5. Monitoring Statistics: A set of summary statistics that serve to monitor the performance of entity resolution systems, even in the absence of a benchmark data set (section 4.3.1).

Our framework does black box evaluation. That is, we evaluate the end result of an entity resolution system, without considering its specific architecture. This allows our framework to apply to any entity resolution system, as long as it produces a clustering as an output.

Figure 4.1 represents the elements of the evaluation framework as well as their interdependencies.

4.1.1 Previous Work

Evaluation is a critical element of iterative model development, model selection, and validation of results. Despite the central importance of evaluation in the development and implementation of entity resolution systems, the topic has received scant attention in the literature.

The most well-studied aspect of evaluation for entity resolution concerns the definition of performance evaluation metrics such as precision and recall (Bilenko and Mooney, 2003), the b-cubed metric (Bagga and Baldwin, 1998), generalized merge distances (Maidasani et al., 2012), and the use of crowdsourcing for data labeling (Christophides et al., 2021). However, this literature does not account for the statistical challenges involved in estimating these metrics from limited data (e.g., a non-representative benchmark data set or a small sample) which is necessary for their reliable use. The statistical literature focused on these estimation challenges

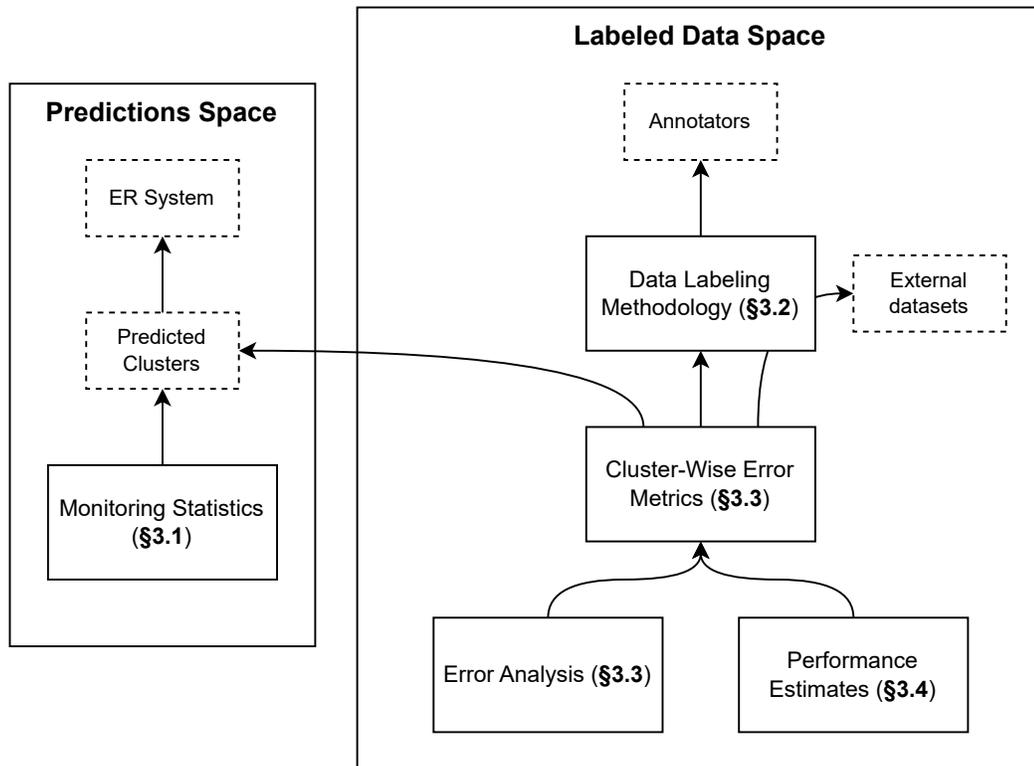


FIGURE 4.1: Diagram representation of the main elements of the framework and their dependencies. The entity resolution system and its predictions live in what we call the “prediction space.” Monitoring statistics can be computed for predictions, while their true value for a ground truth clustering can be estimated using labeled data. In the labeled data space, cluster-wise error metrics are obtained from external benchmark datasets or from our data labeling methodology. Error analysis and performance estimates rely on cluster-wise error metrics.

is also limited (Marchant and Rubinstein, 2017; Dasylyva et al., 2020; Binette et al., 2023). Furthermore, the topics of quality assurance, monitoring, and error analysis appear mostly understudied in the entity resolution literature.

Typical evaluation procedures used in entity resolution can result in misleading conclusions, as some commonly-used performance estimators are biased (Wang et al., 2022; Binette et al., 2023). This can lead to inaccurate representation of model performance and erroneous ranking of competing methods. In particular, the naive computation of pairwise precision on benchmark data sets has been shown to provide

over-optimistic results (Binette et al., 2023). Precision computed on benchmark data sets is often close to 1, even when the true precision for the entire data set may be much lower. When combining biased precision estimates with recall estimates into an F1 score, this can lead to performance rank reversals: an algorithm may be assessed to perform better than another with high confidence, despite the opposite being true in practice (Binette et al., 2023). In other words, comparing entity resolution algorithms based on F1 scores computed on benchmark data sets (such as in Yin et al. (2020)) leads to rankings not representative of performance on larger populations.

4.1.2 Outline of the Paper

The rest of the paper is structured as follows. In section 4.2, we provide background on our motivating application and current approaches to evaluation. In section 4.3, we introduce the proposed methodology. In section 4.4, we showcase its application to inventor name disambiguation and its validation in simulation studies. Finally, we conclude in section 4.5 with a summary of our contributions and directions for future work.

4.2 Background

We begin this section with an overview of the disambiguation work carried out by the American Institutes for Research (AIR) for PatentsView.org. We then discuss our motivating data and application in more detail. Finally, we provide additional information on current industry standards for the evaluation of entity resolution systems, using the methodology used by Statistics Canada as an example.

4.2.1 Patent Data Disambiguation for PatentsView.org (♠)

PatentsView is a public patent data platform maintained by the American Institutes for Research and the U.S. Patents and Trademarks Office. It increases the value,

utility, and transparency of U.S. patent data by providing enriched data products, data visualizations, and data exploration tools.

As one of its main contributions, PatentsView disambiguates patent inventors, assignees, lawyers, and locations. This embeds patent data in a large knowledge graph that links these individual entities through co-authorship, ownership, and citation relationships. However, this disambiguation is a significant challenge given the absence of unique identifiers for these entities, and given the large amount of noise and ambiguity in the data. PatentsView addresses this challenge by employing an assortment of disambiguation algorithms and updating the disambiguation for new data on a quarterly basis.

A variety and growing set of users have made use of PatentsView’s data since its full launch in January 2017. Around 50,000 users visited the PatentsView website in 2023, with 75,000 downloads of bulk data files and an average of one million API requests per month. In Figure 4.2, we show the estimated numbers of citations to PatentsView in the academic literature over time and by Dewey Decimal subject classification.

4.2.2 Motivating Data and Application

We consider inventor data from U.S. patents granted between 1976 and November 2023, inclusively, obtained from the bulk data download page of PatentsView.org (USPTO, 2023). These data are aggregated in a single table indexed by inventor mentions, where each row of this table is referred to as a record. An inventor mention is a reference to a specific inventor (identified by authorship sequence number) on a specific patent (identified by patent number). For each inventor mention, we have information such as inventor first and last name, inventor location, patent co-inventor names, patent title, patent abstract, patent application date, patent grant date, patent assignee, and patent classification codes; see Table 4.1 for examples.

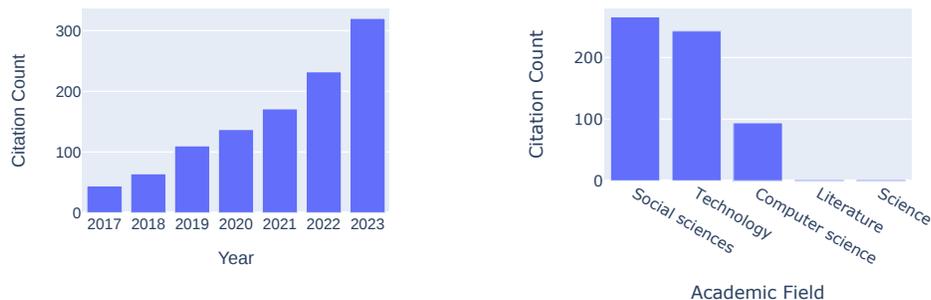


FIGURE 4.2: **Left:** Estimated number of citations to PatentsView in academic literature by year. **Right:** Number of citations by estimated Dewey Decimal Broad Classification. The estimated citation numbers were obtained by searching Google Scholar for mentions to “PatentsView” and “Patents View” and reviewing all results, with the 2023 year estimate containing extrapolated counts for November and December. The Dewey Decimal Classification categories were obtained by extracting abstracts from papers and programmatically querying openAI’s GPT-3.5 model for a classification estimate. Note that GPT-3.5 could not ascertain the classification code for 189 papers. (♠)

Additionally, we have the results of PatentsView’s disambiguation algorithm applied to 18 individual data releases between August 2017 and November 2023. At each release, more inventor mentions were disambiguated as new patents were granted. Furthermore, the disambiguation algorithm was occasionally tweaked in this time period. For each data release, a disambiguation is available that consists of a membership vector assigning each inventor mention to a unique inventor identifier. Our evaluation framework specifically targets this sequence of disambiguation results.

Table 4.1: Example of attributes available for individual inventor mentions. Additional attributes, such as co-inventor information and patent text, are available as well.

Mention ID	Patent	Name	City	State	Country	Year	Title	Kind	Assignees	Classification
US8501339-3	8501339	Lutgard C. De Jonghe	Lafayette	CA	US	2013	Protected lit...	B2	['PolyPlus Battery Company']	H01G
US6085742-1	6085742	Stuart Lindsay	Tempe	AZ	US	2000	Intrapulmonary del...	A	['Aeromax Technologies, Inc.']	A61M
US9215286-6	9215286	Elin R. Pedersen	Portola Valley	CA	US	2015	Creating a social	B1	['Goolge Inc']	H04L

A number of benchmark data sets are available to help assess the performance of PatentsView’s disambiguation system (Binette, 2022). Most of these data sets provide “ground truth” disambiguation (of various quality) of a specific subset of inventors for a given time period. They are all cluster samples that fit our evaluation framework. However, except for the new inventor benchmark data developed in (Binette et al., 2023; Binette, 2022), none of the benchmark data sets are representative of the full population of inventors or have sampling weights associated with them.

4.2.3 Industry Standards for Entity Resolution Evaluation (♠)

We look at Statistics Canada to understand common industry practices for the evaluation of entity resolution systems. The institution has been at the forefront of record linkage since the seminal paper of Fellegi and Sunter (1969), who were then working at Statistics Canada, and with Fellegi becoming Chief Statistian of Canada from 1985 to 2008. Research groups at Statistics Canada have continued to advance the field in recent years, including on the topic of evaluation.

For context, Statistics Canada routinely performs the linkage of many data sets including censuses, administrative data, and survey data. These linkages are classified according to their purpose as analytical or operational. In the former case, the main goal is to produce a linked file including all the required responses and explanatory variables, to fit some regression model. A good example is the linkage of the Canadian Community Health Survey to the Canadian mortality database (Sanmartin et al., 2016). Many of the analytical linkages are implemented within the Social Data Linkage Environment (Canada, 2022). All other linkages are classified as operational, such as linkages that support specific steps in a sample survey, e.g., frame maintenance operations, like removing duplicates or linking different sampling frames in a multi-frame survey. For all the linkages, regardless of their purpose, the accuracy of the linkage decisions is measured with clerical reviews or a statistical model.

When the purpose is analytical, a linked data set is created that may end up in a research data center, including quality indicators for the data users. According to Qian et al. (2021), the indicators should comprise the linkage rate, the linkage representativeness (LR) according to van der Laan and Bakker (2015), and the precision and the false negative rate to characterize the linkage accuracy at the level of the record pairs. The LR metric highlights potential biases, here linkage rate differences across subgroups, by measuring the variance of estimated linkage propensities.

At Statistics Canada, accuracy measures are estimated primarily through clerical reviews (Dasylyva et al., 2016) that can be carried out with G-LINK, the agency's generalized system for probabilistic record linkage. These reviews consist of visual inspections on a probability sample of record pairs to determine if their constituent records refer to the same unit. When the linkage is probabilistic, the pairs are typically stratified according to their linkage weights (Millard and Blanchard, 2022, chap. 5). For quality control, the same pair may be reviewed by two or more clerks (Dasylyva et al., 2016). For analysis, further activities are conducted to evaluate the linkage quality (Statistics Canada, 2017, chap. 6, chap. 7.5).

For operational linkages, the linkage accuracy may also be measured with clerical reviews as in the census over-coverage study (Statistics Canada, 2019, chap. 8.2.1), where the reviews are based on sampling groups of connected records (Dasylyva et al., 2015). These groups are essentially connected components in the graph where the vertices are the records and the edges are the links. To cut costs, Statistics Canada is also considering model-based estimates (Dasylyva and Goussanou, 2022). For example, this approach was used to set the record similarity threshold (i.e., how similar two records have to be before they are linked) in the probabilistic linkage between the 2021 census of agriculture and the census of population for the same year (Canada, 2023).

4.3 Methodology

We now describe our evaluation framework in detail.

Let \mathcal{R} be a set of N records. Each record in \mathcal{R} is a reference to a unique entity (an entity mention), with some entities being referred to by multiple records. Two records are said to be coreferent, or to match if they refer to the same entity. Note that coreference is an equivalence relation: if records A and B refer to the same entity, and records B and C refer to the same entity, then records A and C also refer to the same entity. As such, coreference induces a clustering \mathcal{C} of \mathcal{R} , with two records being in the same cluster if they refer to the same entity. We refer to \mathcal{C} as the “true” clustering, i.e., it represents the true identity relations that entity resolution aims to recover. An entity resolution system outputs a predicted clustering $\hat{\mathcal{C}}$. To evaluate the entity resolution system, we introduce a methodology that samples clusters $c \in \mathcal{C}$ and use them to assess the accuracy of $\hat{\mathcal{C}}$. Throughout, we use the $c(r)$ to denote the cluster in \mathcal{C} containing a given record $r \in \mathcal{R}$. Similarly, we let $\hat{c}(r)$ be the predicted cluster in $\hat{\mathcal{C}}$ containing a given $r \in \mathcal{R}$.

To fix ideas, we can take \mathcal{R} as the set of inventor mentions on granted U.S. patents. The clustering \mathcal{C} represents the true grouping of inventor mentions according to the real-world inventor that they represent, i.e., each cluster $c \in \mathcal{C}$ links to a set of patents authored by this inventor. The predicted clustering $\hat{\mathcal{C}}$ is the output of an inventor disambiguation system.

4.3.1 Summary Statistics and Quality Assurance

The first component of our evaluation framework is a set of summary statistics that describe properties of any given disambiguation result. The goal of these statistics is to provide key indicators that can be tracked to understand and monitor disambiguation results throughout the lifetime of an entity resolution system. They

are simple and easily interpretable statistics that can help explain properties of the clustering. Additionally, these statistics act as quality assurance indicators that can be automatically monitored to identify potential bugs and errors. In particular, estimated summary statistics can be monitored using standard quality assurance tools such as control charts (Montgomery, 2020) and anomaly detection (Chandola et al., 2009).

The statistics we propose are described below.

Definition 3 (Cluster Size Distribution Statistics). Let \mathcal{C} be a clustering of a set of records \mathcal{R} . We define the following metrics regarding the distribution of cluster sizes in \mathcal{C} .

Average Cluster Size: The average cluster size is defined as $R_{\text{size}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} |c|$.

Matching Rate: The matching rate statistic R_m is the proportion of records $r \in \mathcal{R}$ that are linked to some other records, i.e., are part of a cluster with at least two records. Namely, $R_m = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathbb{I}(|c(r)| > 1)$.

Cluster Hill Numbers (Entropy Curve): For a given order $q \geq 0$, the corresponding Hill number H_q is the exponentiation of the Rényi entropy of order q of the cluster size distribution. That is, given a clustering \mathcal{C} , we have

$$H_q = \left(\sum_{i=1}^N (\mathbb{P}(|c| = i))^q \right)^{1/(1-q)}, \quad (4.1)$$

where $\mathbb{P}(|c| = i) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbb{I}(|c| = i)$. Here, (4.1) is continuously extended at $q = 0$ and $q = 1$, and taking the limit from the left for $q = \infty$.

The set of Hill numbers $\{H_q\}_{q \geq 0}$ uniquely characterizes the ordered cluster size distribution and provides interpretable statistics. For instance, H_0 is the number of unique cluster sizes in the distribution; H_1 is the exponential Shannon entropy;

H_2 is the inverse of the probability that two random inventors have authored the same number of patents; and, H_∞ is the prevalence of the most common cluster size. Since Hill numbers are continuous in their parameter q , they provide a simple representation of the ordered cluster size distribution as a continuous curve.

The next set of metrics quantifies the level of noise in the data. Suppose that each cluster element r is associated with a label, such as an inventor’s name listed on a patent. For a given inventor, listed names may differ on different patent applications. Additionally, multiple inventors may share the same name. To quantify these two situations, we introduce the homonymy rate and name variation rate statistics.

Definition 4 (Variation and Homonymy Rate Statistics). Let \mathcal{C} be a clustering of a set of records \mathcal{R} , where each record is associated with a label s_r , and let $n(r)$ be the set of records with the same labels as r , i.e., $n(r) = \{r' \in \mathcal{R} \mid s_{r'} = s_r\}$. We define the following metrics to describe label similarity across clusters and label variation within clusters.

Homonymy Rate: The homonymy rate R_h is the proportion of clusters containing a record that shares its label with another cluster. That is, the homonymy rate is defined as

$$R_h = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbb{I}(\exists r \in c : n(r) \not\subseteq c(r)). \quad (4.2)$$

Name Variation Rate: The name variation rate R_v is the proportion of clusters with variation among the record labels. That is, the name variation rate is defined as

$$R_v = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbb{I}(\exists r \in c : c(r) \not\subseteq n(r)). \quad (4.3)$$

Remark 8. Various measures of cluster homogeneity/compactness and separation have been proposed in the literature and are used as part of the objective functions

of unsupervised clustering algorithms (Johnson, 1967; Davies and Bouldin, 1979; Liu et al., 2010; Duran and Odell, 2013). When using clustering algorithms relying on such statistics, these can be reported in addition to our proposed summary statistics.

The above summary statistics can either be applied to the predicted clustering $\hat{\mathcal{C}}$ (replacing \mathcal{C} by $\hat{\mathcal{C}}$ in definitions 3 and 4), or estimated for the unknown ground truth clustering \mathcal{C} by using a sample of clusters. That is, suppose that a sample of k clusters, c_1, \dots, c_k , is taken from the true clustering \mathcal{C} . We write p_c for the probability (up to a normalizing constant) that a given cluster c is sampled in any one draw. For example, with probability proportional to size sampling, we have $p_c = |c|/N$. Sampling processes and designs are discussed in more detail in section 4.3.2.

To facilitate estimation of the average cluster size and matching rate in Definition 3, we re-express R_{size} and R_m as ratios of expectations. In particular, suppose we randomly sample a single cluster c , where the probability of sampling a given $c \in \mathcal{C}$ is proportional to $p_c > 0$. The quantities in the numerators and denominators of R_{size} and R_m can be rewritten as expectations with respect to this sampling distribution. That is,

$$R_{size} = \frac{\mathbb{E}[|c|/p_c]}{\mathbb{E}[1/p_c]}, \quad R_m = \frac{\mathbb{E}[|c|\mathbb{I}(|c| > 1)/p_c]}{\mathbb{E}[|c|/p_c]}. \quad (4.4)$$

For example, when sampling clusters with probability proportional to size, we have $p_c = |c|/N$. Consequently, $\mathbb{E}[1/p_c] = \sum_{c \in \mathcal{C}} [1/p_c] p_c = |\mathcal{C}|$, which matches the denominator of R_{size} in Definition 3. We then use the sample of k clusters to estimate the expectations in the numerators and denominators separately—for example, we can estimate each expectation with its corresponding weighted sample average—and take the ratio of the relevant estimated expectations. Alternatively, as described in section 4.3.4.2, we also can use the bias-adjusted estimator and variance estimator described in Binette et al. (2023). Similarly, the homonymy rate and name variation rate can be written as ratios of expectations corresponding to the numerators and denominators

of (4.2) and (4.3), which then can be estimated in the same way.

Estimating the Hill numbers is a more challenging task. Several proposals for estimating a whole curve of Hill numbers exist in ecological and biological diversity estimation (Chao, 1984; Chao et al., 2013, 2014). These apply in our context only when considering uniform sampling weights. We therefore leave the problem of estimating Hill numbers for future work.

4.3.2 Data Labeling Methodology

Data labeling is often needed to construct benchmark data sets suitable for evaluation. Here, we expand and formalize the methodology introduced in Binette et al. (2023) for practical and cost-effective data labeling.

Fundamentally, our goal is to obtain a probability sample of known, ground truth clusters. To do so, we suggest sampling k records r_1, \dots, r_k and having data annotators recover the associated clusters $c(r_1), \dots, c(r_k)$.

A variety of tools can help labelers build these ground truth clusters, including using predicted clusters or blocking as a starting point, and using search tools for identifying candidate matches. Below, we suggest a simple methodology that we have found useful in our application.

The methodology relies on two main components, namely (a) a predicted disambiguation used as the starting point of the data labeling, and (b) a search tool used to identify candidate matches for manual review. For (a), we use an entity resolution algorithm to provide a set of predicted clusters to aid in the disambiguation. In the application to PatentsView.org described in Binette et al. (2023), the current inventor disambiguation was used as the starting point of data labeling. Otherwise, a simple exact name-matching disambiguation could be used. For (b), we used PatentsView.org’s search tool. It is also possible to use spreadsheets to parse through subsets of records, or to use elasticsearch (Elastic, 2022) as a search backend. The use

of elasticsearch provides tolerance to typographical errors and intuitive term-based search functionality that may be familiar to data labelers.

Given a population of records \mathcal{R} and a sample size k , the data labeling methodology is as follows:

- (i) First, sample a sequence of records $S = (r_1, r_2, \dots, r_k)$, $r_i \in \mathcal{R}$.
- (ii) For each sampled record r , we recover the corresponding predicted cluster $\hat{c}(r)$ from (a) above. We then perform the following three steps:
 - (A) The data labeler takes note of overclustering errors: records in the predicted cluster $\hat{c}(r)$ that are not part of $c(r)$. We denote by A_r the set of overclustering errors that the data labeler aims to identify, defined as

$$A_r = \hat{c}(r) \setminus c(r). \quad (4.5)$$

- (B) The data labeler takes note of underclustering errors: records that are in $c(r)$ but that are not in $\hat{c}(r)$. These can be found by using the search tool (per (b) above). We denote by B_r this set of underclustering errors that the data labeler aims to identify, defined as

$$B_r = c(r) \setminus \hat{c}(r). \quad (4.6)$$

- (C) Given A_r and B_r , the true cluster associated with record r is $c(r) = \hat{c}(r) \setminus A_r \cup B_r$.

At the end of this process, we have a sequence of sampled ground truth clusters $c(r_1), c(r_2), \dots, c(r_k)$ associated with each sampled record. Put together, the ground truth clusters form a benchmark data set C_S that can be used for evaluation.

4.3.2.1 Sampling Schemes

Many different designs can be used to sample clusters. We recommend randomly sampling records with replacement and finding associated clusters, which is straightforward to implement and facilitates estimation. In this case, the probability that a given cluster c is sampled in any single draw is $|c|/N$, i.e., this is sampling clusters with probability proportional to their sizes.

Sampling clusters with probability proportional to $|c|$ can result in increased accuracy relative to simple random sampling of clusters. In particular, when a cluster-level outcome of interest is correlated with the size of the cluster, the probability proportional to size design offers smaller standard errors in estimates of population quantities (Lohr, 2021). This is the case for PatentsView data, as large clusters tend to be associated with increased chances for errors.

Other sampling designs can be leveraged, in which case the probabilities p_c should change to match the design. For example, uniform sampling probabilities $p_c \propto 1$ can be appropriate when entities, rather than records, are sampled at random. As another example, suppose the disambiguation algorithm provides match probabilities between all pairs of records; that is, we have probabilities $p_{r,r'}$ that records $r, r' \in \mathcal{R}$ are a match for all (r, r') . Then, for a given record $r \in \mathcal{R}$, the expected number of records to be removed to the predicted cluster $\hat{c}(r)$ in step (A) of the data labeling methodology, the expected overclustering error from (4.5), is

$$\mathbb{E}[|A_r|] = \sum_{r' \in \hat{c}(r)} (1 - p_{r,r'}). \quad (4.7)$$

Similarly, the expected number of records to be added to the predicted cluster $\hat{c}(r)$ in step (B) of the data labeling methodology, the expected underclustering error from (4.6), is

$$\mathbb{E}[|B_r|] = \sum_{r' \in \mathcal{R} \setminus \hat{c}(r)} p_{r,r'}. \quad (4.8)$$

We can sample records with probabilities proportional to the sum of these two expectations, which could facilitate more accurate estimation of key metrics. However, computing p_c would be more complicated in this design, as it requires fitting a probabilistic record linkage model as a first step.

4.3.2.2 Quality Control

Following data labeling, a quality control step is used to identify obvious errors in the labeling. Through automated methods and validation with the data labeler, this helps correct typographical errors and unintentional errors without changing the intent of the labeler.

The first two properties used for quality control are the fact that $A_r \subset \hat{c}(r)$ and that $r \notin A_r$. If the set of overclustering errors identified by a data labeler does not satisfy these constraints, then an error was made. To help identify errors in the set of underclustering errors identified by a data labeler, one can look for records that are not part of the same block as r , or that have highly dissimilar attributes to r . These simple validations can catch most typographical and annotation errors in our experience.

4.3.3 Error Analysis

We now consider the problem of analyzing entity resolution errors identified through data labeling. This is done in complement to performance estimation (section 4.3.4) that provides representative performance metrics, such as precision and recall, for a given population of records \mathcal{R} .

To motivate error analysis, note that adequate performance as measured by performance metrics is a necessary but insufficient characteristic of machine learning systems (Zhang et al., 2022). Complex or black-box machine learning systems can fail in intricate or unexpected ways that are not acceptable, even when good overall

performance is achieved (Oakden-Rayner et al., 2020). For example, systematic failure to disambiguate inventor names from a given culture, for instance due to differences in naming conventions, could be considered an unacceptable flaw even if the relative prevalence of such inventors is low. It is therefore necessary to test systems and to investigate errors to help identify such issues (Zhang et al., 2022; Poth et al., 2020). Broadly, the goals of error analysis are to:

- (i) Identify patterns in the error space. This includes identifying areas of low performance and performance disparities between subgroups.
- (ii) Identify systematic failures and their cause. For example, a simple systematic failure may be related to the use of punctuation marks in names.

Our approach to error analysis has two main steps. In section 4.3.3.1, we define record-wise and cluster-wise error metrics to obtain an interpretable and relevant error space to analyze. In section 4.3.3.2, we show how to analyze performance disparities by subgroups and we perform error auditing to identify common causes for errors. Note that many other approaches from the machine learning testing literature could also be relevant but go beyond the scope of this paper (Murphy et al., 2008; Ramanathan et al., 2016; Zhang et al., 2022; Braiek and Khomh, 2020; Aggarwal et al., 2019; Tuncali et al., 2020).

4.3.3.1 Error Metrics Defined at the Record and Cluster Levels

We propose metrics to quantify the errors made by a predicted clustering $\hat{\mathcal{C}}$, defined as follows.

Definition 5 (Record-Wise Error Metrics). Let \mathcal{C} be a clustering of a set of records \mathcal{R} , let $\hat{\mathcal{C}}$ be a predicted clustering of \mathcal{R} , and let $r \in \mathcal{R}$ be a given record. We define the following error metrics for comparing the true entity cluster $c(r)$ associated with r to the predicted cluster $\hat{c}(r) \in \hat{\mathcal{C}}$ associated with r :

Error Indicator (EI): This is a binary error indicator defined as $\text{EI}(r) = 0$ when the predicted cluster $\hat{c}(r)$ equals the true cluster $c(r)$, and $\text{EI}(r) = 1$ otherwise.

Size Difference Error (SDE): This is the difference in size between the predicted cluster $\hat{c}(r)$ and the true cluster $c(r)$, defined as $\text{SDE}(r) = |\hat{c}(r)| - |c(r)|$.

Overclustering Error (OCE): This is the number of records in the predicted cluster $\hat{c}(r)$ that are not part of the true cluster $c(r)$, defined as $\text{OCE}(r) = |A_r| = |\hat{c}(r) \setminus c(r)|$.

Underclustering Error (UCE(r)): This is the number of records in the predicted cluster $\hat{c}(r)$ that are not part of the true cluster $c(r)$, defined as $\text{UCE}(r) = |B_r| = |c(r) \setminus \hat{c}(r)|$.

Additionally, we define the relative overclustering error (ROCE) as $\text{ROCE}(r) = \text{OCE}(r)/|\hat{c}(r)|$ and the relative underclustering error (RUCE) as $\text{RUCE}(r) = \text{UCE}(r)/|c(r)|$.

For a given cluster $c \in \mathcal{C}$, we can define corresponding metrics by averaging over its internal records $r \in c$.

Definition 6 (Cluster-Wise Error Metrics). Given a cluster c and a record-wise error metric \mathbf{E} , we extend \mathbf{E} to clusters by defining $\mathbf{E}(c) = \frac{1}{|c|} \sum_{r \in c} \mathbf{E}(r)$, the average of record-wise error metrics within the cluster. Specifically, we define the cluster-wise error metrics

$$\text{OCE}(c) = \frac{1}{|c|} \sum_{r \in c} \text{OCE}(r), \quad \text{UCE}(c) = \frac{1}{|c|} \sum_{r \in c} \text{UCE}(r), \quad \text{SDE}(c) = \frac{1}{|c|} \sum_{r \in c} \text{SDE}(r),$$

and

$$\text{ROCE}(c) = \frac{1}{|c|} \sum_{r \in c} \text{ROCE}(r), \quad \text{RUCE}(c) = \frac{1}{|c|} \sum_{r \in c} \text{RUCE}(r), \quad \text{EI}(c) = \frac{1}{|c|} \sum_{r \in c} \text{EI}(r).$$

4.3.3.2 Error Auditing and Statistical Analyses

To understand errors and their causes, we consider cluster-wise error metrics, looking at the characteristics of individual errors, and analyzing their relationship with features of interest. We choose to focus on errors at the cluster level rather than at the record level since this can be more interpretable in some applications, including for inventor disambiguation tasks. Here, the clusters represent individual inventors, and errors for disambiguating a given inventor can be understood in view of inventor characteristics.

To illustrate, for PatentsView’s inventor disambiguation, we consider two analyses based on cluster-wise error metrics. First, we consider marginal performance disparity between imputed inventors’ ethnicities by computing performance metric estimators (see section 4.3.4) within subgroups. We visualize the results by adapting the performance bias module of the Deepchecks Python package (Chorev et al., 2022) to include uncertainty quantification. Here, ethnicity is imputed using the Ethnicolr Python package (Laohaprapanon et al., 2022) with a model trained on the 2010 Census Surname Files. These types of models have inaccuracies, but they can help uncover failure modes that would otherwise remain hidden (Jain et al., 2022).

Second, to audit and classify errors, we manually review the list of errors, identify meaningful categories, and report error rates across the categories (see Figure 4.7). The manual review is assisted by a Streamlit web app (Streamlit, 2023) to browse disambiguated clusters, visualize the differences between predicted and ground truth clusters, and log error tags (see Figure 4.3). Specifically, for a given disambiguated inventor, the user is shown a scatterplot of inventor mentions organized by membership to predicted clusters on the vertical axis versus true clusters on the horizontal axis. All inventor mentions associated with predicted clusters that intersect the true cluster are shown, ensuring that both overclustering and underclustering errors can be visualized.

Hovering over an inventor mention’s data point shows related information, including stated name, assignee, location, patent title, patent grant date, and co-author last names. Additionally, a raw data table can be explored, sorted, and searched, in order to analyze errors in more details. For each disambiguated inventor in a review sample, overclustering and underclustering errors are tagged if present according to a potential cause for error.

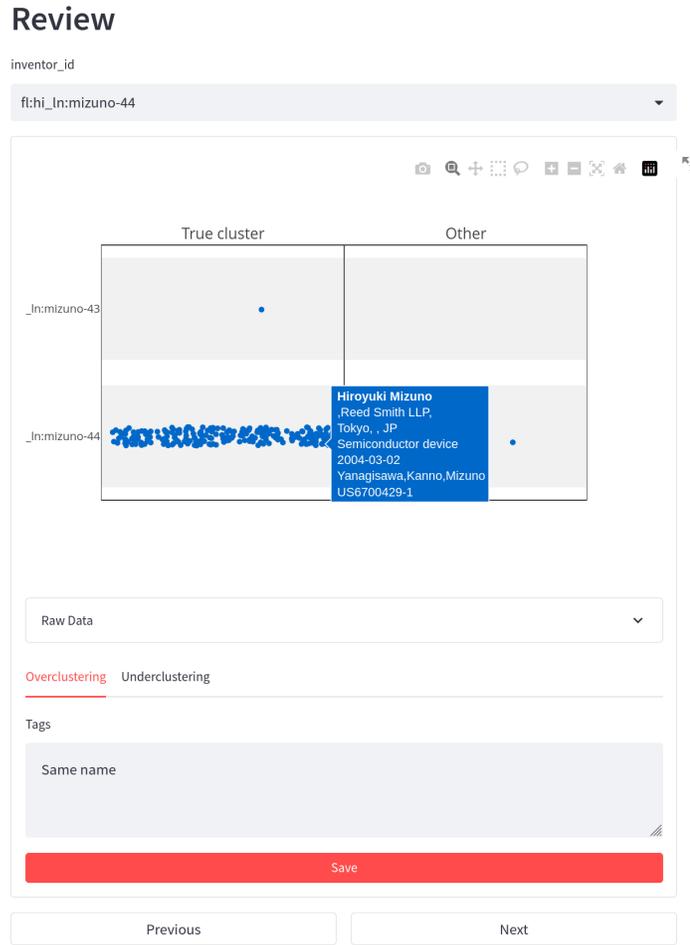


FIGURE 4.3: Screenshot of the Streamlit app used for clerical error review. The “inventor_id” field at the top of the page selects a ground truth cluster whose label is derived from the predicted cluster used as a starting point. The table below shows how this ground truth cluster (first horizontal column) relates to predicted clusters it intersects with on the x axis, with each point representing an inventor mention. Observations regarding overclustering and underclustering errors are recorded below.

Remark 9. Additional error analysis techniques can be relevant in some applications. For instance, decision tree classification models can be used to relate cluster error metrics with cluster features, as done in the SliceFinder algorithm (Chung et al., 2019). The ER-Evaluation Python package (Binette and Reiter, 2023) implements decision tree fitting and visualization tools for this purpose.

Remark 10. Defining performance metrics for subgroups requires some care. Suppose that a subset of records $\mathcal{R}' \subset \mathcal{R}$ is associated with a group of entities of interest, such as inventors of a given ethnicity. Naively, it may be tempting to restrict both $\hat{\mathcal{C}}$ and \mathcal{C} to \mathcal{R}' , before computing cluster-wise error metrics and estimators. This is not the correct approach, as this is blind to errors in predicted links between records in \mathcal{R}' and records not in \mathcal{R}' , and will artificially inflate performance metrics estimates. The correct approach is to first identify entities $\mathcal{C}' \subset \mathcal{C}$ corresponding to the subgroup of interest. Then, cluster-wise error metrics can be computed for sampled clusters that fall within \mathcal{C}' , and performance estimates can be obtained by restricting the sample to this subset. For example, with pairwise precision, this corresponds to estimating the ratio of the number of pairwise links within clusters in \mathcal{C}' (true links) to the number of links in $\hat{\mathcal{C}}$ that intersect \mathcal{C}' (predicted links involving \mathcal{C}').

4.3.4 Performance Metric Estimation

We now turn to the problem of estimating performance evaluation metrics based on benchmark (i.e., labeled) data sets. We assume that the benchmark data sets take the form of a probability sample of true clusters $C_S = (c_1, \dots, c_k)$. We denote by $p_c, c \in \mathcal{C}$, the per-instance sampling probabilities, up to a global normalizing constant.

Many performance metrics commonly used for entity resolution (pairwise, b-cubed, and cluster metrics) can be expressed in terms of the overclustering (OCE) and underclustering (UCE) error metrics defined in (4.5), (4.6), and section 4.3.3, together with functions of the predicted clustering $\hat{\mathcal{C}}$. This representation has the advantage

of directly relating the data labeling process from section 4.3.2 to performance evaluation metrics. As such, labeling uncertainty can be propagated to the estimation of performance evaluation metrics. Furthermore, this representation disaggregates performance evaluation metrics in terms of cluster-level performance, allowing fine-grained error analysis as shown in section 4.3.3. Finally, this representation provides a unified framework for performance estimation in terms of cluster-wise error rates, allowing for efficient computation of all metrics and estimators from a single table containing overclustering and underclustering error metrics. The framework can be extended to the estimation of additional metrics through the use or specification of appropriate record-level error metrics, as shown in section 4.3.4.3.

4.3.4.1 Representation Lemmas

We now provide the expressions for performance metrics that we use to derive estimators.

4.3.4.1.1 Pairwise Precision and Recall

Let \mathcal{P} be the set of pairs of elements belonging to the same cluster in $\hat{\mathcal{C}}$ (predicted pairs) and let \mathcal{T} be the set of pairs of elements belonging to the same cluster in \mathcal{C} (true pairs). Precision P and recall R are defined as

$$P = \frac{|\mathcal{T} \cap \mathcal{P}|}{|\mathcal{P}|}, \quad R = \frac{|\mathcal{T} \cap \mathcal{P}|}{|\mathcal{T}|}. \quad (4.9)$$

Lemma 4 expresses precision and recall as ratios involving the error metrics defined in section 4.3.3.1.

Lemma 4. Suppose we sample one cluster c from \mathcal{C} at random. Let $p_c > 0$ be proportional to its sampling probability. Then

$$P = \frac{\mathbb{E} [|c|(|c| - 1 - \text{UCE}(c))/p_c]}{\mathbb{E} [|c|(|c| - 1 + \text{SDE}(c))/p_c]}, \quad R = \frac{\mathbb{E} [|c|(|c| - 1 - \text{UCE}(c))/p_c]}{\mathbb{E} [|c|(|c| - 1)/p_c]}. \quad (4.10)$$

4.3.4.1.2 Pairwise F-Score

Let F_β , $\beta > 0$, be the weighted harmonic mean between precision and recall, namely

$$F_\beta = \left(\frac{P^{-1} + \beta^2 R^{-1}}{1 + \beta^2} \right)^{-1}. \quad (4.11)$$

Lemma 5 provides an expression for F_β in terms of our error metrics.

Lemma 5. Suppose we sample one cluster c from \mathcal{C} at random. Let $p_c > 0$ be proportional to its sampling probability. Then

$$F_\beta = \frac{\mathbb{E} [|c| (|c| - 1 - \text{UCE}(c)) / p_c]}{\mathbb{E} \left[|c| \left(|c| - 1 + \frac{1}{1+\beta^2} \text{SDE}(c) \right) / p_c \right]}. \quad (4.12)$$

4.3.4.1.3 Cluster Precision and Recall

Following Menestrina et al. (2010), we define cluster precision cP , cluster recall cR , and cluster F -score as

$$cP = \frac{|\mathcal{C} \cap \hat{\mathcal{C}}|}{|\hat{\mathcal{C}}|}, \quad cR = \frac{|\mathcal{C} \cap \hat{\mathcal{C}}|}{|\mathcal{C}|}, \quad cF_\beta = \left(\frac{cP^{-1} + \beta^2 cR^{-1}}{1 + \beta^2} \right)^{-1}. \quad (4.13)$$

That is, cP is the proportion of correctly predicted clusters among all predicted clusters, and cR is the proportion of correctly predicted clusters among all true clusters.

Lemma 6. Suppose we sample one cluster c from \mathcal{C} at random. Let $p_c > 0$ be proportional to its sampling probability. Then

$$cP = \frac{N \mathbb{E} [\text{EI}(c) / p_c]}{|\hat{\mathcal{C}}| \mathbb{E} [|c| / p_c]}, \quad cR = \frac{\mathbb{E} [\text{EI}(c) / p_c]}{\mathbb{E} [1 / p_c]}, \quad cF_\beta = \frac{\mathbb{E} [N(1 + \beta^2) \text{EI}(c) / p_c]}{\mathbb{E} [(N\beta^2 + |\hat{\mathcal{C}}| |c|) / p_c]}. \quad (4.14)$$

4.3.4.1.4 B-Cubed Precision and Recall

The b-cubed (or B^3) precision and recall (Bagga and Baldwin, 1998), here placing equal weight on each ground truth cluster, are defined as

$$P_{B^3} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|c|} \sum_{r \in c} \frac{|c(r) \cap \hat{c}(r)|}{|\hat{c}(r)|}, \quad R_{B^3} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|c|} \sum_{r \in c} \frac{|c(r) \cap \hat{c}(r)|}{|c(r)|}. \quad (4.15)$$

Lemma 7 expresses the b-cubed metrics in terms of the relative expected numbers of missing or extraneous links.

Lemma 7. Suppose we sample one cluster c from \mathcal{C} at random. Let $p_c > 0$ be proportional to its sampling probability. Then

$$P_{B^3} = \frac{\mathbb{E}[(1 - \text{ROCE}(c))/p_c]}{\mathbb{E}[1/p_c]}, \quad R_{B^3} = \frac{\mathbb{E}[(1 - \text{RUCE}(c))/p_c]}{\mathbb{E}[1/p_c]}. \quad (4.16)$$

4.3.4.2 Performance Estimators

All of the expressions in section 4.3.4.1 are of the form,

$$\theta = \mathbb{E}[f(c)]/\mathbb{E}[g(c)], \quad (4.17)$$

for two functions f and g , and where the expectations are taken with respect to sampling the random cluster c . Under the assumption that c_1, \dots, c_k are sampled with replacement, we can estimate $\mathbb{E}[f(c)]$ and $\mathbb{E}[g(c)]$ using the empirical averages \bar{f}_k and \bar{g}_k , respectively, where

$$\bar{f}_k = \frac{1}{k} \sum_{i=1}^k f(c_i), \quad \bar{g}_k = \frac{1}{k} \sum_{i=1}^k g(c_i). \quad (4.18)$$

We take the ratio of these averages, and thus obtain a ratio estimator for θ . We make further adjustments to reduce bias in the ratio estimator and its variance estimator using the approach described in Binette et al. (2023). That is, our estimate

of quantities of the form (4.17), given samples clusters c_1, \dots, c_k , is

$$\hat{\theta} = \frac{\bar{f}_k}{\bar{g}_k} \left\{ 1 + \frac{1}{k(k-1)} \sum_{i=1}^k \frac{g(c_i)}{\bar{g}_k} \left(\frac{f(c_i)}{\bar{f}_k} - \frac{g(c_i)}{\bar{g}_k} \right) \right\}. \quad (4.19)$$

Our variance estimate is

$$\widehat{V}(\hat{\theta}) = \left(\frac{\bar{f}_k}{\bar{g}_k} \right)^2 \frac{1}{k(k-1)} \sum_{i=1}^k \left(\frac{g(c_i)}{\bar{g}_k} - \frac{f(c_i)}{\bar{f}_k} \right)^2. \quad (4.20)$$

4.3.4.3 Example Extension to an Additional Metric

Our framework can be extended to estimate additional metrics. For example, consider the cluster homogeneity metric, defined as the normalized conditional entropy between the true and predicted clusterings (Rosenberg and Hirschberg, 2007). That is, homogeneity is defined as

$$h = 1 - \frac{H(\mathcal{C} \mid \hat{\mathcal{C}})}{H(\mathcal{C})}, \quad (4.21)$$

where

$$H(\mathcal{C} \mid \hat{\mathcal{C}}) = - \sum_{c \in \mathcal{C}} \sum_{\hat{c} \in \hat{\mathcal{C}}} \frac{|c \cap \hat{c}|}{N} \log \frac{|c \cap \hat{c}|}{|\hat{c}|} \quad \text{and} \quad H(\mathcal{C}) = - \sum_{c \in \mathcal{C}} \frac{|c|}{N} \log \frac{|c|}{N}. \quad (4.22)$$

Define the record-wise error metric \mathbb{H} as

$$\mathbb{H}(r) = (|\hat{c}(r)| - \text{OCE}(r)) \log \frac{|\hat{c}(r)| - \text{OCE}(r)}{|\hat{c}(r)|} \quad (4.23)$$

and, for a cluster $c \in \mathcal{C}$, define the cluster-wise variant $\mathbb{H}(c) = \frac{1}{|c|} \sum_{r \in c} \mathbb{H}(r)$.

Lemma 8. Suppose we sample one cluster c from \mathcal{C} at random with probability proportional to positive numbers $p_c > 0$, $c \in \mathcal{C}$. Then

$$h = 1 - \frac{\mathbb{E}[|c| \mathbb{H}(c) / p_c]}{\mathbb{E}[|c| \log(|c|/N) / p_c]}. \quad (4.24)$$

Similarly as before, expression (4.24) can be used to define a ratio estimator of cluster homogeneity.

4.4 Empirical Illustrations and Simulations

In this section, we first showcase the application our evaluation framework to PatentsView’s inventor disambiguations. We then present results of a simulation study to assess the accuracy of our performance metric estimators.

Our data labeling was performed using the methodology described in section 4.3.2, resulting in a set of 400 cluster samples representative of data up to December 31, 2022. This benchmark data set is a direct extension of the work of Binette et al. (2023), where some practical details of the data labeling process are explained in more detail.

4.4.1 Summary Statistics and Quality Assurance

Figure 4.4 displays our summary statistics computed using PatentsView’s predicted inventor disambiguations $\hat{\mathcal{C}}$ as a function of time. For Hill numbers, we focus on H_0 , the number of distinct cluster sizes, and H_1 , the exponentiated Shannon entropy. Figure 4.4 also displays our estimates of the summary statistics for \mathcal{C} , excluding H_0 and H_1 , for disambiguations carried out on or before December 31, 2021 (black dotted line).¹

Figure 4.4 reveals several features of the evolution of these summary statistics and their estimates. First, the average cluster size, matching rate, and name variation rate statistics computed with $\hat{\mathcal{C}}$ jump quite significantly around 2021, suggesting something unusual has happened in that time frame. Second, the number of distinct cluster sizes is roughly monotonic, except for early disambiguation history and again around 2021. The monotonic trend is in line with expectations, as the number of distinct cluster sizes should increase as data are added over time; the sudden break in the trend around 2021 is not. The homonymy rate statistic drops over time, going down to

¹ We cannot estimate the true value of summary statistics values for later disambiguations since our benchmark data set only covers records up to December 31, 2021.

Disambiguation Summary Statistics

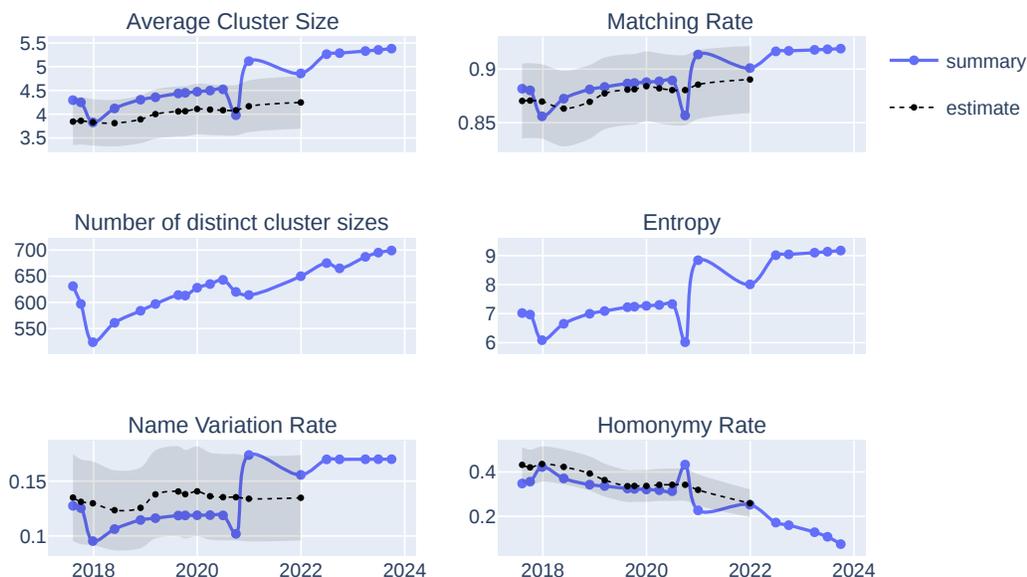


FIGURE 4.4: **Blue line:** Summary statistics for PatentsView’s history of predicted disambiguations. **Black dotted line:** Estimates of the true value of the summary statistics, based on the 2022 inventors’ benchmark data, with pointwise 95% confidence intervals.

nearly 5% by 2024, meaning that almost 95% of inventor’s names are assumed to be unique in the predicted disambiguation. The noticeable changes around 2021 coincide with a change to the disambiguation algorithm, apparently one that impacts the properties of the clusterings. We observe that summary statistics from $\hat{\mathcal{C}}$ are mostly within the confidence intervals for the corresponding quantities in \mathcal{C} . This suggests that the disambiguation algorithm generates clusterings with similar properties (as measured by these summary statistics) as the true clustering. Regardless, the rather significant changes in 2021 should motivate further investigation to ensure that the data still meet quality expectations.

One challenge with interpreting Figure 4.4 is that both the data and algorithm change over time. It is possible to separate these two aspects by considering the evolution of summary statistics for a fixed subset of the data. In Figure 4.5, we consider inventor mentions from before August 2017 as a fixed data set over time. This is the largest data subset that was disambiguated at all available time points, allowing us to see the evolution of summary statistics over PatentsView’s entire history.

The change patterns observed in Figure 4.4 are accentuated in Figure 4.5. For inventor mentions dating from before August 2017, the average cluster size and the homonymy rate from \hat{C} now fall outside of the 95% confidence intervals, even though that was quite not the case in Figure 4.4. Evidently, the quality of the pre-2017 data disambiguation has been affected by changes to the algorithm over the period. In fact, these changes were made to account for the significant amount of new data incorporated in the years between 2017 and 2021. This observation highlights the importance of considering the effect of change both in algorithms and in amount of data when assessing disambiguation quality. Indeed, as we have previously noted, the difficulty of entity resolution problems is not constant across data sizes; rather, the opportunity for errors grows quadratically as a function of data size. Changes made to account for growing data, and specifically to rebalance precision and recall, since false match errors increase the fastest, will necessarily affect the characteristics of the disambiguation of data subsets.

Overall, we recommend using summary statistics as a monitoring and quality control tool, tracking both global behavior and properties of fixed data subsets. Unexpected behaviors, sudden changes, or incompatibility with representative estimates, should trigger an investigation to validate the quality of the system’s inputs and outputs.

Disambiguation Summary Statistics

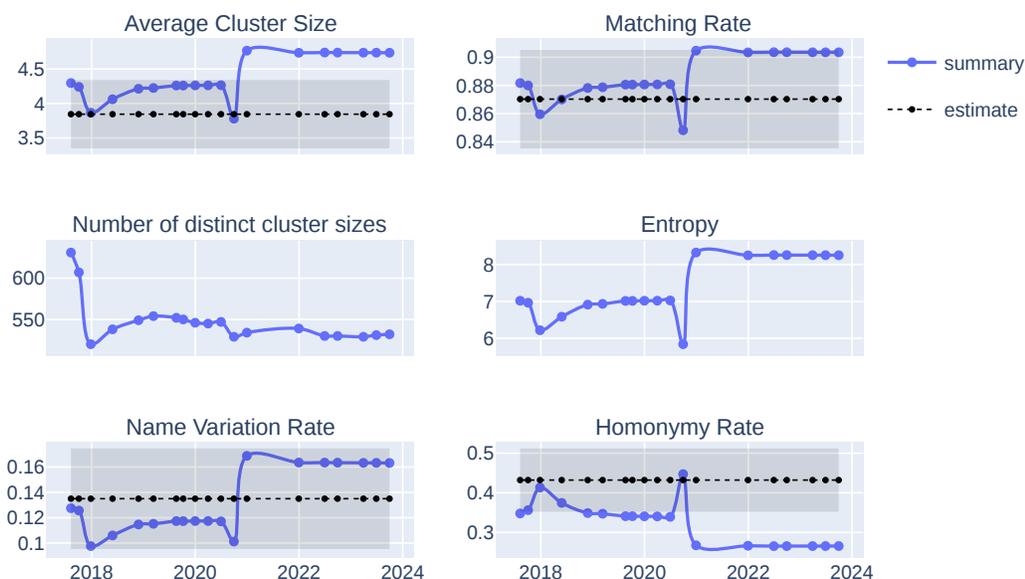


FIGURE 4.5: Summary statistics and estimates for the fixed data set of inventor mentions dating up to August 2017. Disambiguations of this fixed data set have changed over time, as changes to the algorithm were made and since information from additional records was used to resolve entities. As before, the dotted line is the estimate of the summary value for the true clustering of the August 2017 inventor mentions. The shaded bands are pointwise 95% confidence intervals. Since the data set is fixed in this case, the estimates are constant over time.

4.4.2 Performance Estimates

Figure 4.6 displays performance estimates over PatentsView’s disambiguation history, with plus or minus one standard deviation confidence intervals. There is an important dip in performance before the beginning of 2021, which was then corrected. Performance in view of these metric estimates has been mostly stable since 2022, which provides some assurance in the quality of the linkages despite the changes in the algorithm.

Performance Estimates

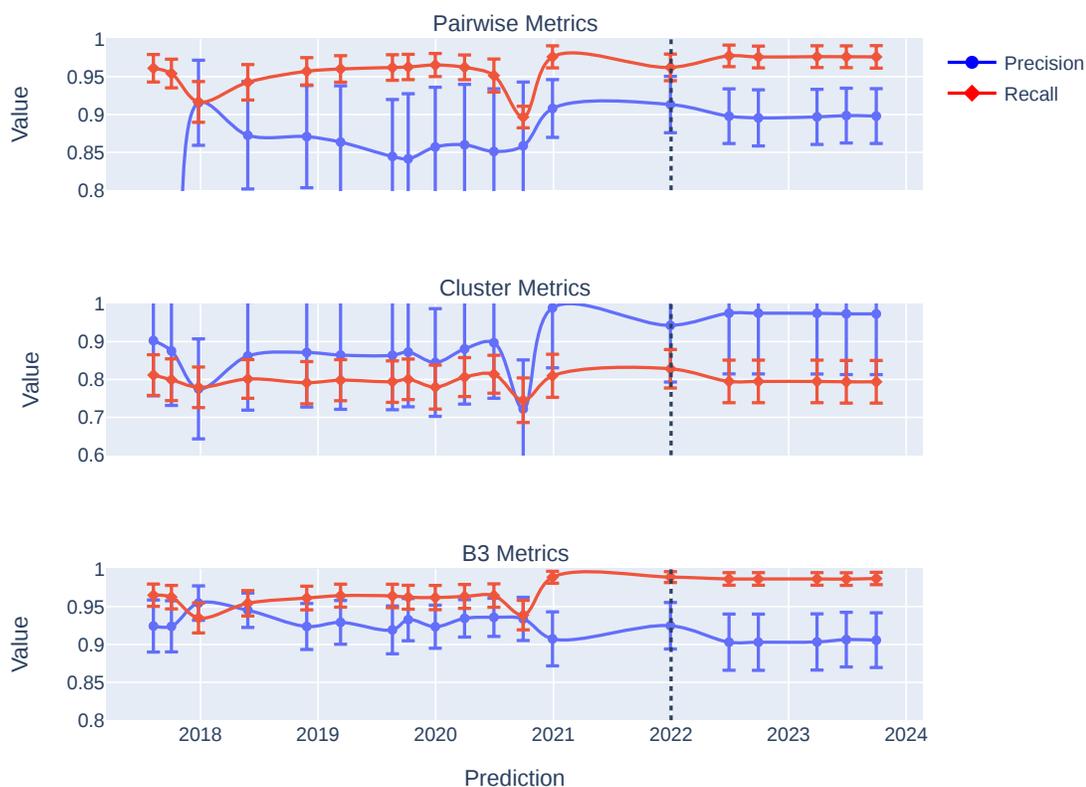


FIGURE 4.6: Performance metrics estimates and confidence intervals (plus or minus one estimated standard deviation) over PatentsView’s disambiguation history. The ground truth data only cover inventor mentions up to December 31, 2021. As such, post-2021 estimates correspond to the accuracy of the disambiguation for data up to that point.

The estimated pairwise precision generally has remained lower than the estimated pairwise recall, and yet estimated cluster precision has been generally higher than estimated cluster recall. There are two things to note about this. First, uncertainty for the cluster precision estimates is large. This is because our sampling scheme, sampling clusters with probability proportional to size, leads to relatively few sampled instances of small clusters, despite small clusters being highly prevalent in the data.

With cluster metrics putting equal weight on all clusters, this leads to higher estimated standard deviation. It would be possible to reduce uncertainty for the cluster metric estimates by increasing the sample size or by specifying an alternate sampling scheme that is more likely to result in the observation of small clusters, for example, by stratifying based on the predicted cluster size corresponding to each record before sampling with probability proportional to size. This could be relatively inexpensive, as manually reviewing small clusters is typically faster than reviewing large clusters.

Furthermore, it is surprising to see pairwise recall estimates being higher than pairwise precision estimates since PatentsView has aimed to provide higher precision than recall (three of our authors have been directly involved in PatentsView). One key application of these metric estimates is to better align the accuracy and characteristics of entity resolution with business objectives. Accurate performance metric estimates can be used as objective functions for training machine learning models, for performing model selection, or for calibrating a given model. Accuracy objectives and the relative balance between metrics can be accounted for to satisfy requirements.

4.4.3 Error analysis

We now turn to the analysis of errors, their causes, and their relationship with features of interest. Figure 4.7 displays the weighted relative frequency of observations made by a clerical reviewer (the first author) when analyzing errors presented using our error auditing tool. In practice, we would want to perform error auditing in two passes. The first would be a brainstorming session, taking notes to identify and define meaningful categories and labels that can be applied to different kinds of errors. A second pass then would implement the strategy derived from the first stage, helping provide actionable insights into causes of errors. Common issues could be investigated further by a development team.

Here, we show the raw data from the first brainstorming step, as it helps illustrate

Clerical Review of Clustering Errors

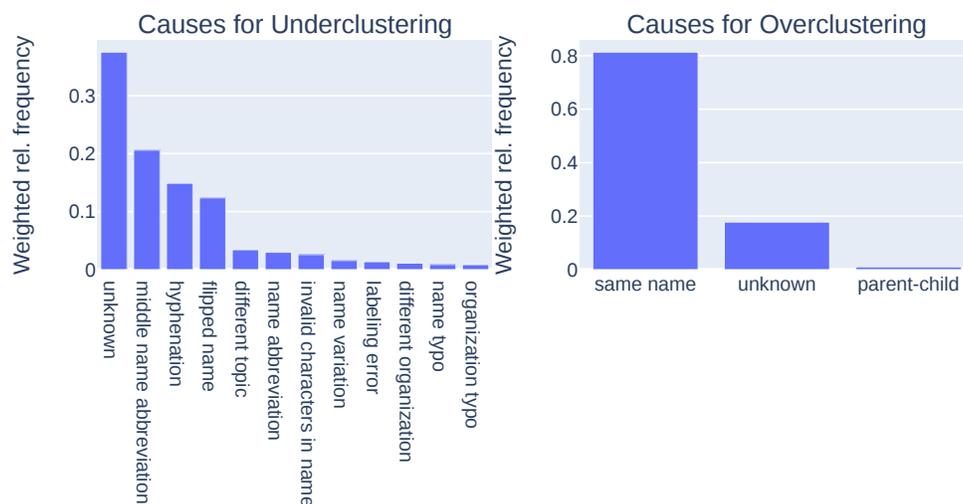


FIGURE 4.7: Reviewer’s notes and their weighted relative frequencies for patterns in overclustering and underclustering errors.

what kinds of observations and insights were first made. One common label is “unknown.” This label is applied when an entity is not correctly disambiguated, but it was not clear what could have been the cause of the error. As examples, a case could be particularly ambiguous; there could be an error in the data labeling; or, there could be insufficient contextual information in the data to justify combining two predicted clusters or separating one predicted cluster into two clusters. A second common label is “same name,” assigned to overclustering errors. This represents cases where inventor mentions were merged because of a shared name, even though other contextual information pointed towards the two representing different inventors. Otherwise, common underclustering errors are associated with variation in name spelling, such as a middle name being abbreviated or not, a name being hyphenated or not, a first and last name being written in one order or the other (which is common in certain cultures), or, less frequently, a typographical error in a name or an invalid

character. A few underclustering errors are associated with the dissimilarity of patent topics or a typographical error in the spelling of the assigned organization. For a few cases, a labeling error might be a cause for the error. Note that the error auditing did not aim at finding errors in the data labeling, and so this label only represents anecdotal observations.

As previously noted, following a first observation and brainstorming step, a precise error auditing plan should be prepared. This plan should include clear definitions of a specific set of labels that can be applied to certain error cases. Following the application of this second step, results regarding key issues in the disambiguation can be communicated and used to consider mitigation methods.

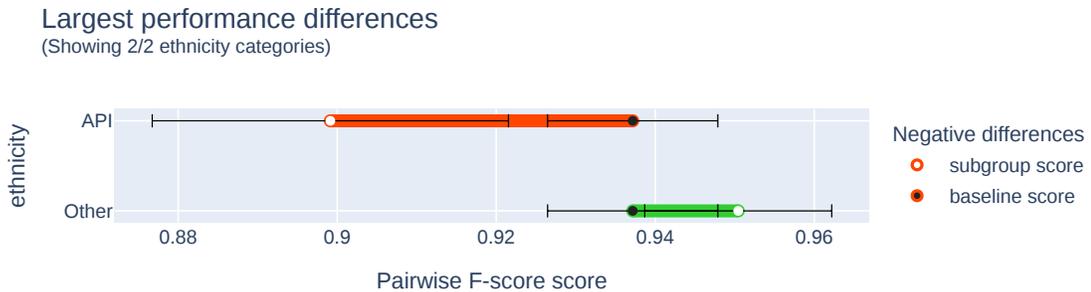


FIGURE 4.8: Performance difference from the baseline for inventors with an inferred Asian and Pacific Islander ethnicity (API) or other inferred ethnicity.

Figure 4.8 displays performance disparities between inventors with an inferred Asian and Pacific Islander ethnicity (API) or other inferred ethnicity. There is an estimated 5% difference in F-score across these two subgroups. Although this is not a very large difference considering the uncertainty in estimation, analyzing performance disparities in this way can help identify subgroups for which more attention should be directed in the design of disambiguation algorithms.

4.4.4 Simulation Study

To conclude this section, we present two simulation studies. The first is based on the RLData10000 data set, and the second second uses PatentsView data up to May 28, 2018.

The RLData10000 simulation is designed to evaluate the effectiveness of sampling with probabilities proportional to cluster size when clusters are small. We use a predicted disambiguation with high pairwise precision (91%) and high pairwise recall (97%), as we believe this makes for a challenging estimation task. Indeed, many clusters will be correctly disambiguated in this case, and therefore few errors will be observed.

The PatentsView simulation is designed to validate the accuracy of our estimators specifically when applied to PatentsView’s data. We use the December 30, 2021, disambiguation as a “ground truth,” as we believe it is a close approximation to the true clustering in terms of cluster size distribution. We use the May 28, 2018 disambiguation as a prediction of the “ground truth”, as it is one of the earliest reliable disambiguations produced by PatentsView. The May 28, 2018, disambiguation has 91% pairwise precision and 94% pairwise recall when compared to the December 30, 2021, disambiguation. This is a high accuracy bar, which makes the performance estimation problem challenging.

4.4.4.1 RLData10000 Simulation

The RLData10000 data set (Sariyar and Borg, 2022) is a synthetic data set containing 10,000 personal information records with first name, last name, birth date, birth month, and birth year. There are 1,000 clusters of size two and 8,000 singleton clusters. Individuals’ names and birth dates sometimes appear with errors, and distinct individuals sometimes share the same names or birth dates. As previously discussed, we consider an “all-but-one” matching algorithm for our predicted disambiguation.

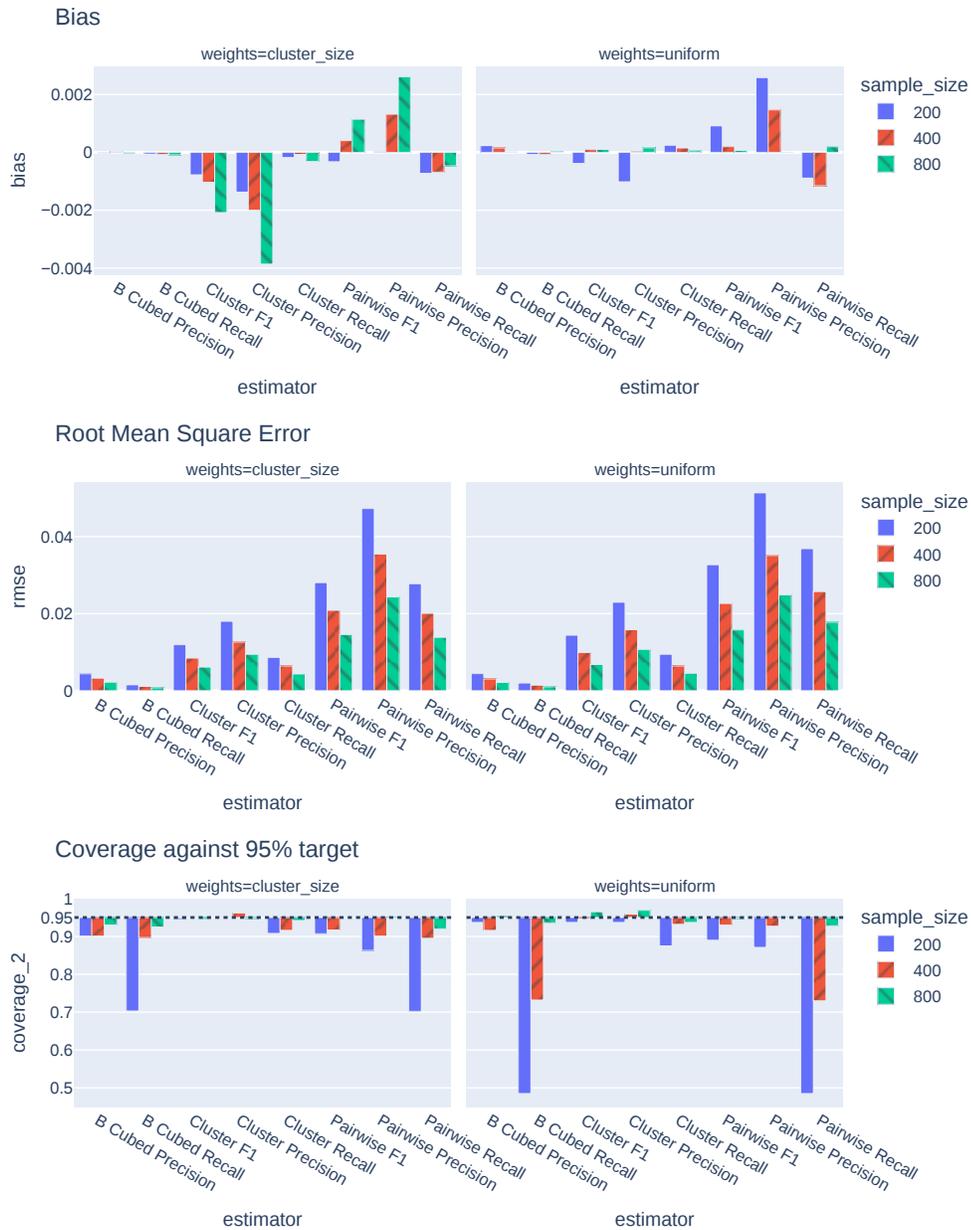


FIGURE 4.9: Simulation study based on the RLdata10000 data set. The accuracy of an “all-but-one” matching algorithm was estimated by sampling ground truth clusters in a simulation replicated 1000 times per set of parameters. The estimates were compared to the known accuracy of the algorithm, specifically, 91% pairwise precision and 97% pairwise recall. Note that bias is below 0.4% in all cases.

That is, we link together two records if and only if they match on four or five components among of first name, last name, birth year, birth month, and birth day.

For evaluation, we sample clusters with replacement, either with probability proportional to cluster size (the default approach) or with uniform probabilities. These designs result in weights labeled respectively as “cluster_size” and “uniform.” We consider samples of sizes 200, 400, and 800. For each combination of parameters and estimator, we replicate the sampling and estimation process 1,000 times. Since we have ground truth for the RLData10000 data set, we can compute the bias and root mean square error (RMSE) of the point estimates. We also compute empirical coverage rates of approximate 95% confidence intervals. We consider coverage since the extent of any deviation from the nominal coverage is easy to visualize and understand. We note that it is generally challenging to achieve good coverage from large-sample confidence intervals, especially when dealing with sparse data or skewed distributions, as the distribution of the estimator is only approximately normal.

Figure 4.9 summarizes the results of 1,000 runs. The empirical bias is always smaller than 0.4% and always less than 0.2% when going up to samples of size at least 400. This validates the near unbiasedness of the estimators.

In terms of RMSE, the pairwise metric estimators are the least accurate, followed by cluster estimators, and then the highly accurate b-cubed estimators. To interpret the RMSE values for the pairwise precision estimator, we first point out that a data-free (and not recommended) estimator of precision equal to 100% achieves RMSE of 9%, since the true pairwise precision is 91%. With probability proportional to cluster size sampling and at a sample of size 200, the RMSE is around 4.7%. This decreases to 3.5% at sample size 400 and 2.4% at sample size 800. These RMSE values, while not insubstantial, tend to be smaller than the corresponding RMSEs from uniform probability sampling. To see why the cluster and b-cubed estimators are more accurate, note that the pairwise estimators are defined in terms of pairs of records that are predicted to match or that are true matches. There are only 1,000 matching pairs of records across the 9,000 clusters in this data set, and there is a

roughly similar number of predicted matching pairs. This makes estimation difficult as a large number of sampled clusters will not be associated with any predicted or matching pair. On the other hand, cluster and b-cubed metrics are defined relative to the populations of true and predicted clusters, for which we collect information in each sample.

Finally, we consider the coverage of the approximate 95% confidence intervals. The “coverage_2” label represents confidence intervals defined as the point estimate plus or minus two times the estimated standard deviation. The coverage rate for both precision and recall estimators is low at sample size 200. However, with sampling probability to cluster size, coverage rates are at least 90% when using samples of size 400 or larger. At sample size 800, the coverage rate is roughly nominal. The coverage rates when sampling with uniform probability weights are lower in general. This is due to the fact that errors are more rarely observed with a uniform design, leading to sparse data and a more variable standard deviation estimator. Overall, we attribute the less-than-nominal coverage rates in lower sampler sampler size to two factors, namely non-normality of the estimator’s sampling distribution and excessive variability in the standard deviation estimates. This is evident in the distribution of the standard deviation estimator, displayed in Figure 4.10 for the pairwise precision and pairwise recall estimators when sampling with probability proportional to cluster size. With size 200, the empirical distribution of the pairwise recall standard deviations has a point mass at 0. This corresponds to cases where no underclustering errors were observed in the sample, leading to a recall estimate of 100% with 0 standard deviation.

Overall, we take away from the simulation study that the estimators can offer accurate reflections of the performance metrics, especially when using probability proportional to size sampling with sufficient sample sizes.

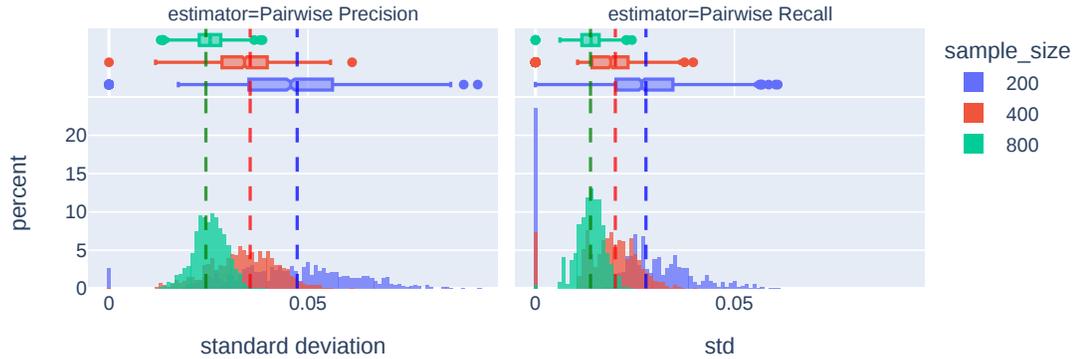


FIGURE 4.10: Distribution of the standard deviation estimator for pairwise precision and pairwise recall estimators, with probability proportional to cluster size sampling.

4.4.4.2 PatentsView Data Simulation

We now consider the simulation based on PatentsView data. We use the same parameters as for the RLData10000 simulation described in section 4.4.4.1. Figure 4.11 displays the results of the simulation.

In terms of bias, uniform sampling weights are unreliable for estimating pairwise precision in this simulation. We do not recommend using uniform sampling weights for data like the PatentsView data, as large cluster sizes with many errors may not be observed frequently in the sample. On the other hand, the bias of estimators is negligible when using sampling with probability proportional to cluster sizes.

Considering RMSE, in this simulation, we see that sampling with probability proportional to cluster size leads to reasonable RMSEs for b-cubed and pairwise metrics estimators, and the uniform sampling design continues to be inadequate for PatentsView data, as evident by the large RMSE of the corresponding estimators. The reason for the large RMSE of cluster metrics estimates, when sampling with probability proportional to cluster size, is the same as for their unreliable confidence

interval coverage discussed in 4.4.2. That is, relatively few small clusters are sampled, despite small clusters being the most prevalent. Since the cluster metrics put equal weight on all clusters, with small clusters being the most prevalent, this leads to increased variability of the estimates.

Considering the coverage of confidence intervals, we see similar behavior as in the RLData10000 simulation when sampling with probability proportional to cluster size. Coverage rates approach the nominal 95% level as the sample size grows. Sample sizes of 400 or 800 appear necessary to obtain reasonable coverage with these data. Confidence intervals of pairwise metrics are unreliable when sampling clusters uniformly.

Overall, the simulation study demonstrates the effectiveness our estimators when using simple random sampling of records, which is a convenient design for implementation and corresponds to sampling clusters with probability proportional to their size. Accurate estimates can be obtained for sample sizes that are practical in real applications.

4.5 Discussion

This paper introduces a novel evaluation framework for entity resolution systems. It ties an entity-centric data labeling methodology together with informative evaluation tasks, such as monitoring, performance estimation, and error analysis. Furthermore, the framework unifies many aspects of the evaluation process through the definition of two key metrics defined at the record or cluster level: the overclustering error and the underclustering error (section 4.3.3.1). All of the performance estimators are derived from these two metrics or simple variants, making it straightforward to relate error analysis with performance metric estimates and to extend the framework to estimate additional performance metrics.

We have demonstrated how our framework can be used in practice, without requiring the use of sophisticated sampling schemes. Once a weighted benchmark data set has been collected using our data labeling methodology, it can be used to evaluate multiple different disambiguation algorithms in multiple different ways. The labeled data are not tied to a singular algorithm or evaluation objective. Furthermore, we validated the estimators and data labeling methodology in simulation studies, providing evidence that sampling clusters with probability proportional to size (via sampling records uniformly at random) can facilitate accurate estimation of key metrics in different situations.

There is opportunity for future work on the design of refined sampling schemes and estimators. A finite population sampling point of view would be useful to accommodate smaller data sets and data sets with very large clusters. Adaptive sampling schemes, or sampling schemes derived from model-based estimates, could improve efficiency. Finally, more sophisticated ratio estimators and variance estimators could be considered and compared, including model-based or model-assisted estimators that use covariates available at the record level. Our unified evaluation framework provides opportunity for such sampling schemes and estimators to be useful for a large range of evaluation objectives. Furthermore, our framework accommodates the propagation of labeling uncertainty into estimates. This is an important topic that can be explored in more depth. Finally, it would be useful to be able to extrapolate the performance of a given algorithm over time, as more records are collected, or when applied to larger datasets. This could help extrapolate performance from artificial benchmarks to real data, or help anticipate performance degradations. Our current evaluation framework can be used as a starting point for these problems.

Software

Our evaluation framework is implemented in the “ER-Evaluation” Python package (Binette and Reiter, 2023) available at <https://github.com/OlivierBinette/er-evaluation/>.

Author Contributions

Olivier Binette led the project, the methodological research and development, and the writing. Youngsoo Baek contributed to the development of summary statistic estimators. Siddharth Engineer provided context and user analytics for PatentsView, using OCR and large language models to classify hundreds of papers citing PatentsView. Christina Jones contributed to the development of methods and provided context and user analytics for PatentsView. Abel Dasylyva contributed the section on industry standards and to the methodology. Jerome P. Reiter contributed to the development of the statistical methods and the writing.

4.6 Appendix

Proofs

Proof of Lemma 4. First, we express $|\mathcal{P}|$ as a sum over $c \in \mathcal{C}$ through

$$|\mathcal{P}| = \sum_{\hat{c} \in \hat{\mathcal{C}}} \binom{|\hat{c}|}{2} = \sum_{r \in \mathcal{R}} \frac{1}{|\hat{c}(r)|} \binom{|\hat{c}(r)|}{2} = \frac{1}{2} \sum_{c \in \mathcal{C}} \sum_{r \in c} (|\hat{c}(r)| - 1). \quad (4.25)$$

For a given cluster $c \in \mathcal{C}$ and $r \in c$, adding and subtracting $|c|$, we find $|\hat{c}(r)| - 1 = |c| - 1 + \text{SDE}(r)$. Substituting this expression into (4.25), we obtain

$$|\mathcal{P}| = \frac{1}{2} \sum_{c \in \mathcal{C}} \sum_{r \in c} (|c| - 1 + \text{SDE}(r)) = \frac{1}{2} \sum_{c \in \mathcal{C}} |c| (|c| - 1 + \text{SDE}(c)). \quad (4.26)$$

Similarly, we express $|\mathcal{T} \cap \mathcal{P}|$ as

$$|\mathcal{T} \cap \mathcal{P}| = \sum_{r \in \mathcal{R}} \frac{1}{|\hat{c}(r) \cap c(r)|} \binom{|\hat{c}(r) \cap c(r)|}{2} = \frac{1}{2} \sum_{r \in \mathcal{R}} (|\hat{c}(r) \cap c(r)| - 1). \quad (4.27)$$

Using the fact that for $r \in c$ we have $|\hat{c}(r) \cap c(r)| = |c| - \text{UCE}(r)$, and averaging over $c \in \mathcal{C}$, we find

$$|\mathcal{T} \cap \mathcal{P}| = \frac{1}{2} \sum_{r \in \mathcal{R}} (|c| - \text{UCE}(r)) = \frac{1}{2} \sum_{c \in \mathcal{C}} |c| (|c| - \text{UCE}(c)). \quad (4.28)$$

Finally, it follows from definition that

$$|\mathcal{T}| = \frac{1}{2} \sum_{c \in \mathcal{C}} |c| (|c| - 1). \quad (4.29)$$

The lemma follows directly from our expressions for $|\mathcal{P}|$, $|\mathcal{P} \cap \mathcal{T}|$, and $|\mathcal{T}|$ after re-expressing them as expectations over a random cluster c distributed with probabilities proportional to $p_c > 0$. \square

Proof of Lemma 5. First write

$$F_\beta = \frac{(1 + \beta^2)|\mathcal{T} \cap \mathcal{P}|}{|\mathcal{P}| + \beta^2|\mathcal{T}|}. \quad (4.30)$$

Substituting (4.26), (4.29), and (4.28) in the above, we obtain

$$F_\beta = \frac{(1 + \beta^2) \sum_{c \in \mathcal{C}} |c| (|c| - \text{UCE}(c))}{\sum_{c \in \mathcal{C}} |c| (|c| - 1 + \text{SDE}(c) + \beta^2(|c| - 1))} = \frac{\sum_{c \in \mathcal{C}} |c| (|c| - \text{UCE}(c))}{\sum_{c \in \mathcal{C}} |c| \left(|c| - 1 + \frac{1}{1 + \beta^2} \text{SDE}(c) \right)}. \quad (4.31)$$

The lemma follows after re-expressing the sums as expectations over a random cluster c distributed with probabilities proportional to $p_c > 0$. \square

Proof of Lemma 6. Write $\tilde{p}_c = p_c / \sum_{c' \in \mathcal{C}} p_{c'}$ for the normalized probability mass function of the random cluster c , and let $w_c = |\mathcal{C}|^{-1} / \tilde{p}_c$ be the ratio of the constant probability mass function to \tilde{p}_c . From the fact that $N = \sum_{c \in \mathcal{C}} |c|$ we can derive

$$|\mathcal{C}| = N/\mathbb{E}[|c|w_c]. \quad (4.32)$$

As such, we find

$$|\mathcal{C} \cap \hat{\mathcal{C}}| = \sum_{c \in \mathcal{C}} \mathbf{EI}(c) = |\mathcal{C}| \mathbb{E}[\mathbf{EI}(c)w_c] = \frac{N\mathbb{E}[\mathbf{EI}(c)w_c]}{\mathbb{E}[|c|w_c]} \quad (4.33)$$

and, after simplifying normalizing constants,

$$cP = \frac{|\mathcal{C} \cap \hat{\mathcal{C}}|}{|\hat{\mathcal{C}}|} = \frac{N\mathbb{E}[\mathbf{EI}(c)/p_c]}{|\hat{\mathcal{C}}|\mathbb{E}[|c|/p_c]}. \quad (4.34)$$

The expression for cR is a standard self-normalized importance sampling representation, i.e.,

$$cR = \frac{|\mathcal{C} \cap \hat{\mathcal{C}}|}{|\mathcal{C}|} = \sum_{c \in \mathcal{C}} |c|^{-1} \mathbf{EI}(c) = \frac{\mathbb{E}[\mathbf{EI}(c)w_c]}{\mathbb{E}[w_c]} = \frac{\mathbb{E}[\mathbf{EI}(c)/p_c]}{\mathbb{E}[1/p_c]}. \quad (4.35)$$

Finally, using (4.32) and (4.33), we find

$$cF_\beta = \frac{(1 + \beta^2)|\mathcal{C} \cap \hat{\mathcal{C}}|}{|\hat{\mathcal{C}}| + \beta^2|\mathcal{C}|} = \frac{(1 + \beta^2)\mathbb{E}[\mathbf{EI}(c)w_c]}{|\hat{\mathcal{C}}|/|\mathcal{C}| + \beta^2} = \frac{(1 + \beta^2)\mathbb{E}[\mathbf{EI}(c)/p_c]}{|\hat{\mathcal{C}}|\mathbb{E}[|c|/p_c]/N + \mathbb{E}[\beta^2/p_c]}. \quad (4.36)$$

□

Proof of Lemma 7. This lemma follows directly from definitions when using self-normalized importance sampling representations as above. □

Proof of Lemma 8. This follows directly from the definitions (4.21), (4.22), and (4.23). □

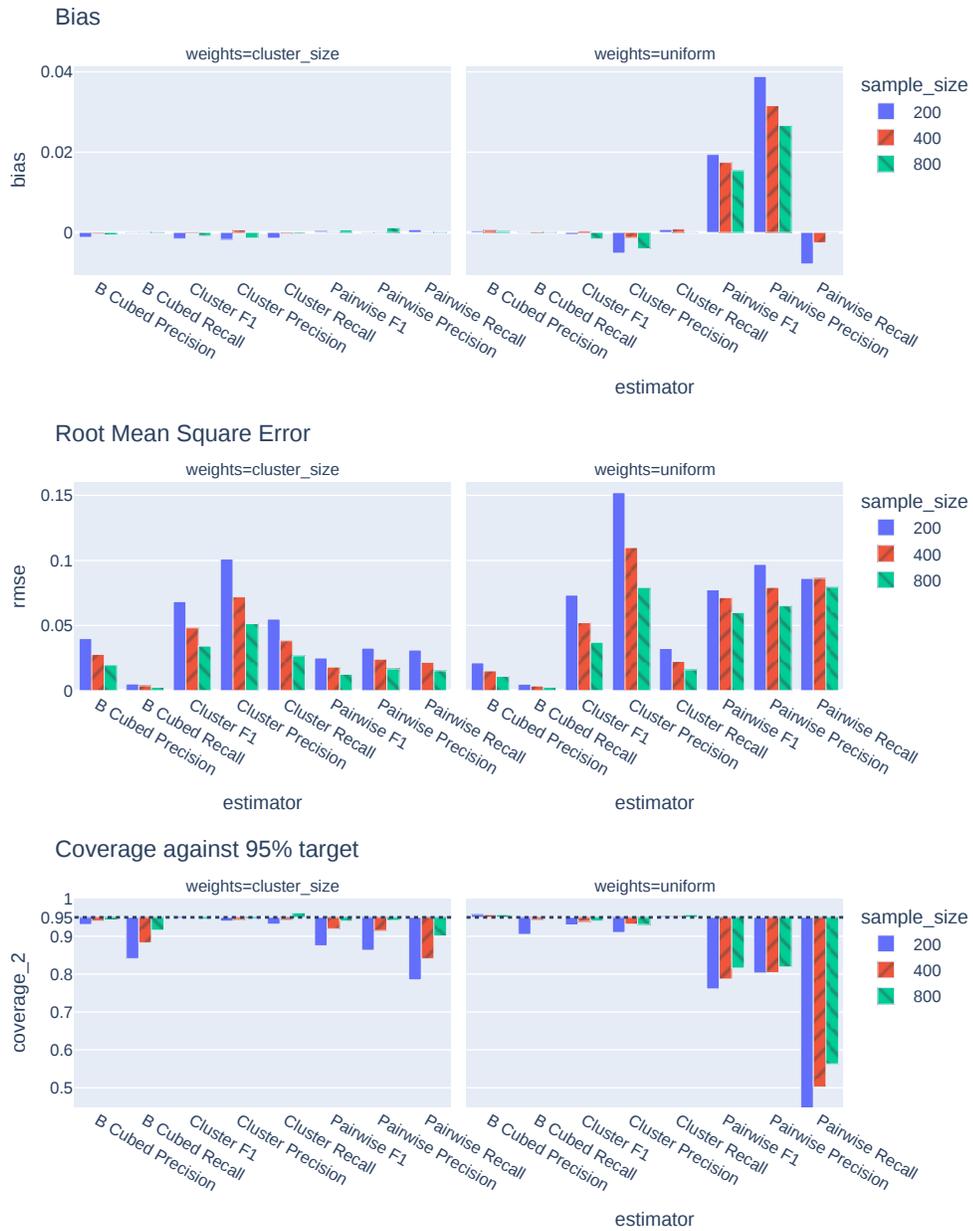


FIGURE 4.11: Simulation study based on the PatentsView’s historical disambiguations, where the December 31, 2021 disambiguation was taken as a prediction, and the May 28, 2018 disambiguation was taken as “ground truth”. The accuracy of the predicted disambiguation was estimated by sampling ground truth clusters in a simulation replicated 1000 times per set of parameters. The estimates were compared to the known accuracy of the algorithm, specifically, 91% pairwise precision and 94% pairwise recall. Note that bias is essentially zero when sampling with probability proportional to cluster size (the default approach), and only substantial for pairwise precision estimates when sampling clusters with uniform probabilities.

5. Optimal F-score Clustering for Bipartite Record Linkage

Probabilistic record linkage is often used to match records from two files, in particular when the variables common to both files comprise imperfectly measured identifiers like names and demographic variables. We consider bipartite record linkage settings in which each entity appears at most once within a file, i.e., there are no duplicates within the files, but some entities appear in both files. In this setting, the analyst desires a point estimate of the linkage structure that matches each record to at most one record from the other file. We propose an approach for obtaining this point estimate by maximizing the expected F -score for the linkage structure. We target the approach for record linkage methods that produce either a posterior distribution of the unknown linkage structure or probabilities of matches for record pairs. Using simulations and applications with genuine data, we illustrate that the F -score estimators can lead to sensible point estimates of the linkage structure.

5.1 Introduction

Often entity resolution (ER) applications involve identifying duplicate records across two databases, where each database has no duplication within. This particular case of ER is called bipartite record linkage, since it corresponds to identifying a bipartite matching between records of the two databases. Bipartite record linkage is used, for example, when matching census records to a post-enumeration survey for

estimating population coverage (Jaro, 1989; Winkler and Thibaudeau, 1991), when linking conflict records for casualty estimation (Sadinle, 2017), in public health and social sciences research (Jutte et al., 2011), and for combining customer records from two lists in a customer relationship management system (Dyché and Levy, 2006).

Bipartite record linkage can be challenging for a variety of reasons, including the computational complexity associated with the comparison of record pairs, bipartite matching and transitivity constraints, and the probabilistic dependencies that these constraints induce in the matching problem. As a result, researchers have proposed a variety of methods for bipartite record linkage, ranging from rule-based approaches to the use of unsupervised clustering models, probabilistic classifiers, and other algorithms supporting different steps of the linkage process (Christophides et al., 2021; Papadakis et al., 2021). For instance, when using the Fellegi and Sunter (1969) framework for probabilistic record linkage to obtain matching weights, one can ensure a bipartite linkage by solving a linear sum assignment problem Jaro (1989). This approach was improved in McVeigh et al. (2019) to incorporate the bipartite linkage constraint directly into a maximum likelihood estimator. Alternatively, one can use a Bayesian model that incorporates the bipartite matching constraint in a prior distribution, resulting in a Bayes linkage estimate under a chosen loss function Sadinle (2017).

Regardless of the approach, it is desirable to have a point estimate of the bipartite linkage structure. However, existing approaches have some limitations. To compute a Bayes estimate, analysts need to specify a loss function, which may be difficult to align with practical requirements. For instance, default parameters of the loss function provided in Sadinle (2017) can lead to a linkage that underestimates the overlap between databases. This causes issues in applications such as casualty estimation, census coverage estimation, and other forms of population size estimation. Similar challenges arise with the use of maximum likelihood estimation, which requires

analysts to tune parameters to estimate the number of matches.

To facilitate the use of bipartite record linkage algorithms, we propose a post-processing step that aligns with a commonly-used objective: to optimize the expected F -score of the resulting linkage under the bipartite record linkage constraint. We consider the F -score, the harmonic mean between precision and recall, as it is arguably the most widely-used evaluation measure in record linkage applications. As a result, optimizing the F -score often aligns with practical requirements.

More precisely, suppose a record linkage algorithm results in pairwise match probabilities or in a posterior distribution on the linkage structure. From those, an expected F -score can be computed for any possible bipartite record linkage. The optimal F -score clustering is the linkage that maximizes this expected F -score. For Bayesian record linkage, this corresponds to a Bayes estimator as described in Section 5.2.2.1. When using pairwise match probabilities, this maximizes a plug-in F -score estimate as described in Section 5.2.2.2.

Our contributions can be summarized as follows.

1. We introduce an efficient algorithm to approximate the optimal F -score under the bipartite record linkage constraint, and use it to construct a point estimate of the linkage structure.
2. We demonstrate that this algorithm can be adapted to any probabilistic bipartite record linkage model.
3. We validate the approach using both simulated and genuine data, demonstrating that the F -score algorithm can yield more accurate point estimates of the linkage structure than some existing point estimators.
4. We discuss accurately estimating the size of the population that overlaps in the two data files.

The remainder of the article is organized as follows. In Section 5.2, we introduce the optimal F-Score algorithm. In Section 5.3, we discuss estimating the overlap population size. In Section 5.4, we investigate the performance of the algorithm using simulated and genuine data. Finally, in Section 5.5, we summarize the findings and discuss potential directions for future research.

5.2 F -score Optimization Under a Bipartite Record Linkage Constraint

Let \mathcal{A} and \mathcal{B} be two data sets comprising $n_{\mathcal{A}}$ and $n_{\mathcal{B}}$ records, respectively. We presume $n_{\mathcal{A}} \geq n_{\mathcal{B}}$, without loss of generality. For $i = 1, \dots, n_{\mathcal{A}}$, each record $\mathcal{A}_i \in \mathcal{A}$ has a unique identifier, denoted \mathcal{A}_{i0} . Similarly, for $j = 1, \dots, n_{\mathcal{B}}$, let \mathcal{B}_{j0} denote the unique identifier for record $\mathcal{B}_j \in \mathcal{B}$. Records \mathcal{A}_i and \mathcal{B}_j are called a match or a link whenever $\mathcal{A}_{i0} = \mathcal{B}_{j0}$, and called a non-match or non-link otherwise. The goal of record linkage is to identify the record pairs $(\mathcal{A}_i, \mathcal{B}_j)$ with $\mathcal{A}_{i0} = \mathcal{B}_{j0}$.

For $i = 1, \dots, n_{\mathcal{A}}$ and $j = 1, \dots, n_{\mathcal{B}}$, let $c_{i,j} = 1$ when \mathcal{A}_i and \mathcal{B}_j are a match, and let $c_{i,j} = 0$ otherwise. Let $\mathbf{C} = [c_{i,j}]$ be the $n_{\mathcal{A}} \times n_{\mathcal{B}}$ matrix with $c_{i,j}$ as the element in the i th row and j th column. We refer to \mathbf{C} as the linkage structure. In bipartite record linkage, we assume that each \mathcal{A}_i is linked to at most one \mathcal{B}_j , and vice versa. That is, for any given indices i and j ,

$$\sum_{j'=1}^{n_{\mathcal{B}}} c_{i,j'} \leq 1, \quad \sum_{i'=1}^{n_{\mathcal{A}}} c_{i',j} \leq 1. \quad (5.1)$$

In practice, \mathbf{C} is an unknown parameter estimated through some record linkage model. Here, we consider record linkage models that result in either a probability distribution for \mathbf{C} , e.g., via a Bayesian model (as in, e.g., Fortini et al., 2001; Tancredi and Liseo, 2011; Gutman et al., 2013; Steorts et al., 2016; Sadinle, 2017; Dalzell and Reiter, 2018; Tang et al., 2020; Betancourt et al., 2022; Guha et al., 2022) or marginal

probabilities $p_{i,j}$ that record pairs $(\mathcal{A}_i, \mathcal{B}_j)$ are matches, e.g., from a Fellegi and Sunter (1969) model as in Enamorado et al. (2019b). With Bayesian record linkage models, the posterior distribution for \mathbf{C} is approximated usually with some Markov chain Monte Carlo sampler, resulting in L plausible draws $\{\mathbf{C}^{(s)} : s = 1, \dots, L\}$ of the linkage structure. These draws can be used to compute Monte Carlo estimates of expectations or other summaries.

The posterior distribution for \mathbf{C} or the estimated match probabilities generally derive from a specified probabilistic model. For now, we do not specify any particular model, since our approach to F -score optimization is agnostic to the structure of the model.

5.2.1 F -score Objective Function and Estimators

For any record pair $(\mathcal{A}_i, \mathcal{B}_j)$, let $\hat{c}_{i,j}$ be an estimate of $c_{i,j}$ and $\hat{\mathbf{C}} = [\hat{c}_{i,j}]$ be the corresponding estimate of \mathbf{C} . For any $\hat{\mathbf{C}}$ and some weight $\beta > 0$, the F -score is defined as

$$F_\beta(\hat{\mathbf{C}}, \mathbf{C}) = \frac{(1 + \beta^2) \sum_{i,j} \hat{c}_{i,j} c_{i,j}}{\beta^2 \sum_{i,j} c_{i,j} + \sum_{i,j} \hat{c}_{i,j}}. \quad (5.2)$$

The expression in (5.2) represents the weighted harmonic mean between precision and recall. When $\beta = 1$, the F -score equally weights precision and recall.

We propose to utilize (5.2) to determine a point estimator $\hat{\mathbf{C}}_{opt}$ of \mathbf{C} . Specifically, we seek the $\hat{\mathbf{C}}$ that maximizes (5.2), that is,

$$\hat{\mathbf{C}}_{opt} = \arg \max_{\hat{\mathbf{C}} \in \mathcal{C}} F_\beta(\hat{\mathbf{C}}, \mathbf{C}), \quad (5.3)$$

where \mathcal{C} is the set of linkage structures satisfying the bipartite linkage condition. Of course, typically $F_\beta(\hat{\mathbf{C}}, \mathbf{C})$ is not observable since it requires knowledge of \mathbf{C} . We therefore must estimate $F_\beta(\hat{\mathbf{C}}, \mathbf{C})$ for any $\hat{\mathbf{C}}$ under consideration.

We consider two approaches for finding $\hat{\mathbf{C}}_{\text{opt}}$. The first method is appropriate for settings where a posterior distribution of \mathbf{C} is available. We use the posterior distribution to approximate (5.2) and subsequently obtain a Bayes estimate from (5.3). The second method is suitable for scenarios where we have point estimates for the probability of a link for each record pair, which we denote as $p_{i,j}$. In this context, we approximate (5.2) using a plug-in approach and obtain a point estimate by solving (5.3). We now describe these two approaches, starting with the first method.

5.2.1.1 Bayes Estimator

When a probability distribution for \mathbf{C} is available, the expectation of (5.2) can serve as a score function that we maximize. This yields the Bayes estimator,

$$\hat{\mathbf{C}}_{\text{Bayes}} = \arg \max_{\hat{\mathbf{C}} \in \mathcal{C}} \mathbb{E} \left[F_{\beta}(\hat{\mathbf{C}}, \mathbf{C}) \right]. \quad (5.4)$$

Here, the expectation is taken with respect to the random variable \mathbf{C} . A closed-form expression for $\hat{\mathbf{C}}_{\text{Bayes}}$ does not exist. In Section 5.2.2, we present an optimization algorithm to approximate the solution to (5.4).

5.2.1.2 Optimal Score Estimator

When estimates $\hat{p}_{i,j}$ for record pairs' $p_{i,j}$ are available, we can obtain an estimator by maximizing a plug-in estimate of the F -score in (5.2). That is, we define the optimal score estimator,

$$\hat{\mathbf{C}}_{\text{OS}} = \arg \max_{\hat{\mathbf{C}} \in \mathcal{C}} \frac{(1 + \beta^2) \sum_{i,j} \hat{c}_{i,j} \hat{p}_{i,j}}{\beta^2 \sum_{i,j} \hat{p}_{i,j} + \sum_{i,j} \hat{c}_{i,j}}. \quad (5.5)$$

As with (5.4), no closed-form expression exists for $\hat{\mathbf{C}}_{\text{OS}}$. However, the algorithm in Section 5.2.2 provides an exact solution to (5.5).

5.2.2 Algorithm for Approximating the F -score

We first present the algorithm for approximating $\hat{\mathbf{C}}_{\text{Bayes}}$ from (5.4). This algorithm is sufficiently general to be adapted for determining $\hat{\mathbf{C}}_{\text{OS}}$ from (5.5), which we discuss subsequently.

5.2.2.1 An Approximation of $\hat{\mathbf{C}}_{\text{Bayes}}$

To approximate $\hat{\mathbf{C}}_{\text{Bayes}}$ from (5.4), we adopt the general framework of outer and inner maximization in Jansche (2007). Speaking broadly, for every possible number of matches k within the range $0 \leq k \leq n_{\mathcal{B}}$, we perform an inner maximization. This step involves approximating the $\hat{\mathbf{C}}$ that optimizes (5.2) for a given k , which we denote as $\hat{\mathbf{C}}_{\text{Bayes}}(k)$. Then, in the outer maximization step, we search across all feasible values of k . The $\hat{\mathbf{C}}_{\text{Bayes}}(k)$ that yields the highest value of (5.2) is the point estimator for \mathbf{C} .

More formally, for a given k , let

$$\hat{\mathbf{C}}_{\text{Bayes}}(k) = \arg \max_{\hat{\mathbf{C}} \in \mathcal{C}, \sum_{i,j} \hat{c}_{i,j} = k} \mathbb{E} \left[F_{\beta}(\hat{\mathbf{C}}, \mathbf{C}) \right] = \arg \max_{\hat{\mathbf{C}} \in \mathcal{C}, \sum_{i,j} \hat{c}_{i,j} = k} \sum_{i,j} \hat{c}_{i,j} \mathbb{E} \left[\frac{(1 + \beta^2)c_{i,j}}{\beta^2 \sum_{i,j} c_{i,j} + k} \right]. \quad (5.6)$$

The optimization expression from (5.6) represents an inner maximization step, which is followed by an outer maximization step

$$\hat{\mathbf{C}}_{\text{Bayes}} = \arg \max_{\hat{\mathbf{C}} \in \{\hat{\mathbf{C}}_{\text{Bayes}}(0), \dots, \hat{\mathbf{C}}_{\text{Bayes}}(n_{\mathcal{B}})\}} \mathbb{E} \left[F_{\beta}(\hat{\mathbf{C}}, \mathbf{C}) \right]. \quad (5.7)$$

For any given k , the inner optimization problem in (5.6) can be solved as a linear sum assignment problem (LSAP), subject to the constraint of k links. In general, LSAP solvers can find unique (bipartite) pairings between elements of two sets to maximize a user-specified total score. The constraint $\sum_{i,j} \hat{c}_{i,j} = k$ can be enforced using various methods, as discussed in Ramshaw and Tarjan (2012). We use a data augmentation approach to incorporate the constraint into the original LSAP problem, as we now explain.

For each record pair $(\mathcal{A}_i, \mathcal{B}_j)$, let $\Delta_{i,j}^{(k)} = \mathbb{E} \left[(1 + \beta^2)c_{i,j}/(\beta^2 \sum_{i,j} c_{i,j} + k) \right]$ represent a score for the pair; we use these scores in the LSAP. Let $\mathbf{\Delta}^{(k)}$ be the $n_{\mathcal{A}} \times n_{\mathcal{B}}$ matrix with entry $\Delta_{i,j}^{(k)}$ in the i th row and j th column. We create an augmented matrix $\tilde{\mathbf{\Delta}}^{(k)}$ of dimension $(n_{\mathcal{A}} + n_{\mathcal{B}} - k) \times n_{\mathcal{B}}$ with elements in the i th row and j th column given by

$$\tilde{\Delta}_{i,j}^{(k)} = \begin{cases} \Delta_{i,j}^{(k)} & \text{if } i \leq n_{\mathcal{A}}, \\ M & \text{otherwise,} \end{cases} \quad (5.8)$$

where

$$M = k(1 + \beta^2)/\beta^2 \geq k \max_{i,j} \Delta_{i,j}^{(k)}. \quad (5.9)$$

Letting $\tilde{n}_{\mathcal{A}} = n_{\mathcal{A}} + n_{\mathcal{B}} - k$, (5.6) is transformed into the optimization problem,

$$\arg \max_{\hat{\mathbf{C}}} \sum_{i=1}^{\tilde{n}_{\mathcal{A}}} \sum_{j=1}^{n_{\mathcal{B}}} \hat{c}_{ij} \tilde{\Delta}_{ij}^{(k)} \quad (5.10)$$

subject to,

$$\sum_{i=1}^{\tilde{n}_{\mathcal{A}}} \hat{c}_{i,j} = 1, j = 1, \dots, n_{\mathcal{B}}, \quad (5.11)$$

$$\sum_{j=1}^{n_{\mathcal{B}}} \hat{c}_{i,j} = 1, i = 1, \dots, \tilde{n}_{\mathcal{A}}. \quad (5.12)$$

This is a variant of the general LSAP formulation. The equality constraints in (5.11) and (5.12) ensure that all $n_{\mathcal{B}}$ elements of \mathcal{B} have unique links to some elements represented by the rows $i \in \{1, \dots, n_{\mathcal{A}}, n_{\mathcal{A}} + 1, \dots, n_{\mathcal{A}} + n_{\mathcal{B}} - k\}$ of $\tilde{\mathbf{\Delta}}^{(k)}$.

The value of M in (5.9) is chosen to be large enough that the solution to (5.10) links all elements in the augmented rows represented by $i \in \{n_{\mathcal{A}} + 1, \dots, n_{\mathcal{A}} + n_{\mathcal{B}} - k\}$ to $n_{\mathcal{B}} - k$ elements in \mathcal{B} , leaving only k elements from the rows $i \in \{1, \dots, n_{\mathcal{A}}\}$ of \mathcal{A} linked to k elements of \mathcal{B} . Specifically, M should be bounded below by k times the largest element of $\Delta^{(k)}$. By setting M this way, any reassignment of matches between

	\mathcal{B}_1	\mathcal{B}_2	\mathcal{B}_3
\mathcal{A}_1	0.1	0.4	0.9
\mathcal{A}_2	0.2	0.5	0.8
\mathcal{A}_3	0.3	0.6	0.7
\mathcal{A}_4	4.0	4.0	4.0

FIGURE 5.1: Example of an augmented matrix $\tilde{\Delta}^{(2)}$ used in the LSAP. The circled links optimize the total score for $k = 2$ matches while respecting the bipartite matching. (♠)

\mathcal{A} and \mathcal{B} does not change the total score by more than M . As such, there will be exactly $n_{\mathcal{B}} - k$ links between elements of \mathcal{B} and the “dummy” elements represented by $i \in \{n_{\mathcal{A}} + 1, \dots, n_{\mathcal{A}} + n_{\mathcal{B}} - k\}$, and the remaining k elements from \mathcal{B} will be matched to elements of \mathcal{A} represented by $i \in \{1, \dots, n_{\mathcal{A}}\}$. By disregarding the $n_{\mathcal{B}} - k$ matches from the augmented rows, we isolate the top k matches. Per the objective function in (5.10), these k matches maximize the score.

To illustrate how the solution to (5.10) solves (5.6), consider a simple example where $n_{\mathcal{A}} = n_{\mathcal{B}} = 3$ and $k = 2$. In this scenario, we construct the augmented 4×3 matrix $\tilde{\Delta}^{(2)}$, as depicted in Figure 5.1. Here, the 3×3 sub-matrix across the first three rows and columns represents $\Delta^{(2)}$. The row labeled \mathcal{A}_4 represents the additional row, where each element is given the score $M = 4.0$. The value of M is selected to exceed $k \max_{i,j} \Delta_{i,j}^{(2)} = 1.8$ per (5.9). The optimal solution to (5.10) includes one link ($n_{\mathcal{B}} - k = 1$) to the “dummy” row \mathcal{A}_4 and two links ($k = 2$) to the records in \mathcal{B} . The augmented LSAP solver from (5.10) guarantees that the resulting estimate $\hat{\mathbf{C}}^{(2)}$ (the circled record pairs in Figure 5.1) achieves the highest score with exactly $k = 2$ links.

Since closed-form expressions for $\Delta_{i,j}^{(k)}$ are not available, we use a Monte Carlo approximation based on samples from the posterior distribution of \mathbf{C} . Using the L

posterior samples $\{\mathbf{C}^{(s)} : s = 1, \dots, L\}$, we estimate each $\Delta_{i,j}^{(k)}$ using

$$\Delta_{i,j}^{(k)} \approx \frac{1}{L} \sum_{s=1}^L \frac{(1 + \beta^2)c_{i,j}^{(s)}}{\beta^2 \sum_{i,j} c_{i,j}^{(s)} + k}. \quad (5.13)$$

These can be efficiently computed for all values of (i, j) and k by using appropriate data structures and exploiting the sparsity of \mathbf{C} . Following Sadinle (2017), suppose that samples from the posterior distribution of \mathbf{C} are represented as $n_{\mathcal{B}} \times 1$ vectors $\mathbf{Z}^{(s)} = (Z_1^{(s)}, \dots, Z_{n_{\mathcal{B}}}^{(s)})$ where any $Z_j^{(s)} = i$ when $c_{i,j}^{(s)} = 1$ and $Z_j^{(s)} = n_{\mathcal{A}} + j$ when $c_{i,j}^{(s)} = 0$. Let $\tilde{\mathbf{Z}}$ be the $n_{\mathcal{B}} \times L$ matrix comprising the columns $[\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(L)}]$; and, let $\mathbf{I} = [I_j^{(s)}]$ be the $n_{\mathcal{B}} \times L$ matrix with elements $I_j^{(s)} = 1$ when $Z_j^{(s)} \leq n_{\mathcal{A}}$, i.e., when record \mathcal{B}_j matches an element of \mathcal{A} in iteration s , and $I_j^{(s)} = 0$ otherwise. From $\tilde{\mathbf{Z}}$, we reconstruct the $n_{\mathcal{A}}n_{\mathcal{B}} \times L$ sparse matrix $\tilde{\mathbf{C}} = [c_{i,j}^{(s)}]$ with $n_{\mathcal{A}}n_{\mathcal{B}}$ rows corresponding to pairs $(\mathcal{A}_i, \mathcal{B}_j)$ and L columns corresponding to the posterior samples. For computational purposes, this matrix is represented internally in coordinate list format (Bates et al., 2023) and computed with time complexity $\mathcal{O}(n_{\mathcal{B}}L)$ by iterating through all elements of $\tilde{\mathbf{Z}}$. Then, for a given k , the matrix $\Delta^{(k)}$ can be vectorized by computing

$$\text{Vec}(\Delta^{(k)}) \approx \tilde{\mathbf{C}} \frac{(1 + \beta^2)/L}{\beta^2 \mathbf{I}^T \mathbf{1} + k}, \quad (5.14)$$

where matrix division is understood to be element-wise. Using the sparse matrix representation of $\tilde{\mathbf{C}}$, the multiplication between $\tilde{\mathbf{C}}$ and the column vector $((1 + \beta^2)/L)/(\beta^2 \mathbf{I}^T \mathbf{1} + k)$ in (5.14) is computed with time complexity $\mathcal{O}(n_{\mathcal{B}}L)$.

5.2.2.2 An Exact Solution to $\hat{\mathbf{C}}_{\text{OS}}$

To compute $\hat{\mathbf{C}}_{\text{OS}}$ in (5.5), we again use the framework in Jansche (2007) and first find the inner maximization solutions $\hat{\mathbf{C}}_{\text{OS}}(k)$ satisfying

$$\hat{\mathbf{C}}_{\text{OS}}(k) = \arg \max_{\hat{\mathbf{C}} \in \mathcal{C}, \sum_{i,j} \hat{c}_{i,j} = k} \sum_{i,j} \hat{c}_{i,j} \frac{(1 + \beta^2) \hat{p}_{i,j}}{\beta^2 \sum_{i,j} \hat{p}_{i,j} + k}. \quad (5.15)$$

We use these estimates to determine the outer maximization solution from

$$\hat{\mathbf{C}}_{\text{OS}} = \arg \max_{\hat{\mathbf{C}} \in \{\hat{\mathbf{C}}_{\text{OS}}(0), \dots, \hat{\mathbf{C}}_{\text{OS}}(n_{\mathcal{B}})\}} F_{\beta}(\hat{\mathbf{C}}, \hat{\mathbf{P}}), \quad (5.16)$$

where $\hat{\mathbf{P}} = [\hat{p}_{ij}]$ is the matrix containing all \hat{p}_{ij} .

The algorithm follows the same steps as the one used to approximate $\hat{\mathbf{C}}_{\text{Bayes}}$, except we replace $\Delta_{i,j}^{(k)}$ with its plug-in estimate, $\hat{\Delta}_{i,j}^{(k)} = ((1 + \beta^2) \hat{p}_{i,j}) / (\beta^2 \sum_{i,j} \hat{p}_{i,j} + k)$. Since no Monte Carlo approximations are used, the algorithm produces an exact solution for (5.5).

5.3 Estimation of Overlap Size (\spadesuit)

An important characteristic of any linkage structure \mathbf{C} is the induced number of matching records, $\sum_{i,j} c_{i,j}$. We refer to this quantity as the overlap size. The overlap size is directly related to the size of the joint population after removing duplicates, which we write as $n_{\mathcal{A},\mathcal{B}} = n_{\mathcal{A}} + n_{\mathcal{B}} - \sum_{i,j} c_{i,j}$. Given an estimate $\hat{\mathbf{C}}$, we define the estimated overlap size as $\sum_{i,j} \hat{c}_{i,j}$; similarly, we can define the estimated population size as $\hat{n}_{\mathcal{A},\mathcal{B}} = n_{\mathcal{A}} + n_{\mathcal{B}} - \sum_{i,j} \hat{c}_{i,j}$.

In some bipartite record linkage applications, one objective is to estimate the overlap size accurately. For example, in as in Wortman (2019), researchers linked a file comprising voters who cast provisional ballots to the official state voter registration file. The overlap size represents the number of voters on the official voter registration file—which includes current and no-longer registered voters—who cast provisional ballots.

This quantity is needed to study the effects of local policies that removed voters from the registration rolls. As another example, in checking the quality of the decennial census, the Census Bureau matches records from the collected decennial population census data to records from a post-enumeration survey. They need to estimate the number of individuals who appear in both data sources, i.e., the overlap size, for use in estimating undercount and overcount (<https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/pes.html>).

Since the F -score equally balances precision and recall (when choosing $\beta = 1$), we can expect the F -score optimal linkage to produce accurate estimates of overlap size. To motivate this, we note that the precision P is defined as the ratio of the number of true positive links, $\sum_{i,j} c_{i,j} \hat{c}_{i,j}$, to the true overlap size, $\sum_{i,j} c_{i,j}$. Similarly, the recall R is defined as the number of true positive links divided by the estimated overlap size, $\sum_{i,j} \hat{c}_{i,j}$. Consequently, when $P = R$ for a specific $\hat{\mathbf{C}}$, the induced overlap size estimate should equal the true overlap size, i.e., $\sum_{i,j} \hat{c}_{i,j} = \sum_{i,j} c_{i,j}$. This property suggests choosing an equal balance between precision and recall in the definition of the F -score to be optimized.

Of course, we should evaluate the potential accuracy of estimators of overlap size computed with the F -score approximations from Section 5.2. Before doing so, however, we provide additional motivation for considering this point estimator over existing methods. Here, we focus on the Bayes estimator introduced by Sadinle (2017). While we do not delve into other methods, we highlight that these methods generally do not directly optimize for accurate overlap size and thus may run into similar issues.

5.3.1 Definition of the BRL Estimator (♠)

We begin by presenting the Bayes estimator in Sadinle (2017), which we call the Beta Record Linkage (BRL) estimator. Using the definition of \mathbf{Z} presented in Section 5.2.2.1,

let $\hat{\mathbf{Z}} = (\hat{Z}_1, \dots, \hat{Z}_{n_{\mathcal{B}}})$ be the vector-representation of the estimated linkage structure $\hat{\mathbf{C}}$. The Bayes estimator for \mathbf{Z} is based on an additive loss $L(\mathbf{Z}, \hat{\mathbf{Z}})$ parameterized by positive constants $(\lambda_{10}, \lambda_{01}, \lambda_{11})$. We have

$$L(\mathbf{Z}, \hat{\mathbf{Z}}) = \sum_{j=1}^{n_{\mathcal{B}}} L_j(Z_j, \hat{Z}_j), \quad (5.17)$$

where

$$L_j(Z_j, \hat{Z}_j) = \begin{cases} 0 & \text{if } Z_j \leq n_{\mathcal{A}}, \\ \lambda_{10} & \text{if } Z_j \leq n_{\mathcal{A}}, \hat{Z}_j = n_{\mathcal{A}} + j, \\ \lambda_{01} & \text{if } Z_j = n_{\mathcal{A}} + j, \hat{Z}_j \leq n_{\mathcal{A}}, \\ \lambda_{11} & \text{if } Z_j, \hat{Z}_j \leq n_{\mathcal{A}}, \hat{Z}_j \neq Z_j. \end{cases}$$

Here, λ_{10} is the loss incurred for any record \mathcal{B}_j from deciding it has no link among \mathcal{A} when in fact it does; λ_{01} is the loss incurred for any record \mathcal{B}_j from deciding it links to some record in \mathcal{A} when in fact its match is not in \mathcal{A} ; and, λ_{11} is the loss incurred for any record \mathcal{B}_j from deciding it is linked to some record \mathcal{A}_i when in fact its match is some other record \mathcal{A}_k , where $k \leq n_{\mathcal{A}}$. The BRL estimator is then

$$\hat{\mathbf{Z}}_{\text{BRL}} = \arg \min_{\hat{\mathbf{Z}}} \mathbb{E}[L(\mathbf{Z}, \hat{\mathbf{Z}})]. \quad (5.18)$$

Suppose that we have the posterior distribution of \mathbf{Z} given the data γ used to match records across \mathcal{A} and \mathcal{B} ; we write this as $p(\mathbf{Z}|\gamma)$. We give an example of γ in Section 5.4.1. The BRL estimator can be computed by solving a standard LSAP. In this context, we find the minimizer and use a matrix for the optimization step interpreted in terms of costs rather than scores. The (i, j) element of the $(n_{\mathcal{A}} + n_{\mathcal{B}}) \times n_{\mathcal{B}}$ cost matrix is

$$w_{ij} = \begin{cases} \lambda_{01}P(Z_j = n_{\mathcal{A}} + j|\gamma) + \lambda_{11}P(Z_j \notin \{i, n_{\mathcal{A}} + j\}|\gamma) & \text{if } i \leq n_{\mathcal{A}} \\ \lambda_{10}P(Z_j \neq n_{\mathcal{A}} + j|\gamma) & \text{if } i = n_{\mathcal{A}} + j \\ \infty & \text{otherwise.} \end{cases} \quad (5.19)$$

When both $0 < \lambda_{10} \leq \lambda_{01}$ and $\lambda_{11} \geq \lambda_{10} + \lambda_{01}$, (5.18) has a closed-form solution Sadinle (2017). Specifically, for $j = 1, \dots, n_{\mathcal{B}}$, we have

$$\hat{Z}_j = \begin{cases} i & \text{if } P(Z_j = i|\gamma) > \frac{\lambda_{01}}{\lambda_{01} + \lambda_{10}} + \frac{\lambda_{11} - \lambda_{01} - \lambda_{10}}{\lambda_{01} + \lambda_{10}} P(Z_j \notin \{i, n_{\mathcal{A}} + j\}|\gamma) \\ n_{\mathcal{A}} + j & \text{otherwise.} \end{cases} \quad (5.20)$$

In this case, a necessary condition for linking the pair $(\mathcal{A}_i, \mathcal{B}_j)$ is that $P(Z_j = i|\gamma) > 0.5$ (Sadinle, 2017). This is apparent when adopting the default parameters in Sadinle (2017), namely $\lambda_{10} = \lambda_{01} = 1$ and $\lambda_{11} = 2$. These default parameter settings are used in both the simulation and analyses of Section 5.4.

5.3.2 Conservative Nature of the BRL Estimator (♠)

With the estimator in (5.20), the LSAP algorithm adopts a conservative approach in declaring links, in that it requires $P(Z_j = i|\gamma) \geq .50$ Sadinle (2017). This reduces the risk of incorrect linkages, but it can introduce bias in the estimation of overlap size under certain linkage scenarios. As an example, consider a task of linking unique individuals in a large data file \mathcal{A} —which has near complete coverage of a population—to participants in a survey data file \mathcal{B} that overlaps with \mathcal{A} . Within \mathcal{A} , suppose exactly two distinct individuals \mathcal{A}_1 and \mathcal{A}_2 share identical names with individual $\mathcal{B}_1 \in \mathcal{B}$. As a result, the posterior probabilities that $Z_1 = 1$ or $Z_j = 2$ are both near (but necessarily below) 0.5, as exemplified in Table 5.1. In this case, it is reasonable to conclude that one of \mathcal{A}_1 or \mathcal{A}_2 is the true link to \mathcal{B}_1 , and therefore to count \mathcal{B}_1 as part of the overlap population. However, if we use (5.20), the BRL estimate fails to declare any links for \mathcal{B}_1 in \mathcal{A} , even though the chance that \mathcal{B}_1 is not linked to any elements in \mathcal{A} is only 2%. Similar issues can arise when more than two records in \mathcal{A} are highly plausible links for some \mathcal{B}_j .

Additionally, (5.19) can produce conservative estimates of links even when the conditions in (5.20) do not apply. This is evident in the following proposition.

Table 5.1: Example of a linkage scenario where BRL selects no match for \mathcal{B}_1 when arguably \mathcal{B}_1 should be counted as part of the overlap population. (\spadesuit)

Index i for \mathcal{A}_i or for non-match	$P(Z_1 = i \gamma)$
1	0.49
2	0.49
3	0.00
4	0.00
5	0.00
$n_{\mathcal{A}} + 1$ (non-match)	0.02

Table 5.2: Example of a general scenario where the LSAP algorithm does not declare a link for \mathcal{B}_j when arguably it should be counted as part of the overlap population. (\spadesuit)

Index i for \mathcal{A}_i or for non-match	$P(Z_j = i \gamma)$
1	0.25
2	0.25
3	0.10
4	0.09
5	0.01
$n_{\mathcal{A}} + j$ (non-match)	0.30

Proposition 3 (Sufficient condition for a non-match (\spadesuit)). *The BRL estimator declares record \mathcal{B}_j a non-link, i.e., $Z_j = n_{\mathcal{A}} + j$, if*

$$(\lambda_{10} + \lambda_{01} - \lambda_{11})P(Z_j = n_{\mathcal{A}} + j|\gamma) \geq \lambda_{10} - \lambda_{11} + \lambda_{11} \max_{i:i \leq n_{\mathcal{A}}} P(Z_j = i|\gamma). \quad (5.21)$$

Proof. Suppose that the bipartite matching condition holds. Then, by assumption in the proposition, for all $i \in \{1, \dots, n_{\mathcal{A}}\}$ we have

$$(\lambda_{10} + \lambda_{01} - \lambda_{11})P(Z_j = n_{\mathcal{A}} + j) \geq \lambda_{10} - \lambda_{11} + \lambda_{11}P(Z_j = i|\gamma). \quad (5.22)$$

This implies that

$$\lambda_{01}P(Z_j = n_{\mathcal{A}} + j) + \lambda_{11}(1 - P(Z_j = i|\gamma) - P(Z_j = n_{\mathcal{A}} + j|\gamma)) \geq \lambda_{10}(1 - P(Z_j = n_{\mathcal{A}} + j|\gamma)). \quad (5.23)$$

As a result, we have

$$\lambda_{01}P(Z_j = n_{\mathcal{A}} + j|\gamma) + \lambda_{11}P(Z_j \notin \{i, n_{\mathcal{A}} + j\}|\gamma) \geq \lambda_{10}P(Z_j \neq n_{\mathcal{A}} + j|\gamma). \quad (5.24)$$

The cost matrix defined by (5.19) implies that the LSAP algorithm will choose $Z_j = n_{\mathcal{A}} + j$ for record \mathcal{B}_j . Note that the converse does not generally hold, as there can be instances where the LSAP solver opts for $Z_j = n_{\mathcal{A}} + j$, yet the condition in (5.21) is not satisfied. \square

Proposition 3 reveals the connection between the posterior probabilities of having no links and of the most likely link candidate in \mathcal{A} . Given specified cost parameters $(\lambda_{10}, \lambda_{01}, \lambda_{11})$, for any \mathcal{B}_j with a sufficiently low maximum link probability $\max_{i:i \leq n_{\mathcal{A}}} P(Z_j = i|\gamma)$, the LSAP algorithm always declares a non-link for record \mathcal{B}_j .

For different choices of fixed cost parameters, we obtain different variations of the sufficient condition. A particularly illuminating result follows under unity cost $(\lambda_{10} = \lambda_{01} = \lambda_{11})$, which we state as a corollary below.

Corollary 2 (Sufficient Condition for non-match under unity costs (\spadesuit)). *Under the unity cost assumption $\lambda_{10} = \lambda_{01} = \lambda_{11}$, a sufficient condition for a non-match for record \mathcal{B}_j is*

$$P(Z_j = n_{\mathcal{A}} + j|\gamma) \geq \max_{i:i \leq n_{\mathcal{A}}} P(Z_j = i|\gamma). \quad (5.25)$$

The sufficient condition in (5.25) assigns a non-link to \mathcal{B}_j when its posterior probability of having no links is largest. This particular decision rule is reasonable when the posterior distribution for Z_j is characterized by large probability mass either at $Z_j = n_{\mathcal{A}} + j$ or $Z_j = i$ for some i . However, if the posterior distribution is multi-modal across a variety of potential links in \mathcal{A} , the sufficient condition may be undesirable. To illustrate, consider the example in Table 5.2. According to (5.25), the LSAP will result in a non-link decision. This results mainly because of significant

uncertainty in the match status, as reflected through a posterior distribution with moderate probability masses placed on the decisions $Z_j = 1$, $Z_j = 2$, and $Z_j = n_{\mathcal{A}} + j$. In a comparison space with multiple instances like the one in Table 5.2, all would be designated non-links by the LSAP. Such behavior may lead to underestimation of overlap size in practice; for example, in Table 5.2, the posterior probability of there being a match, $1 - P(Z_j = n_{\mathcal{A}} + j | \gamma)$, is quite high at 70%.

5.4 Simulation Studies and Illustrative Examples

We now evaluate the Bayes estimator based on the F -score using simulation studies and applications. As a comparison, we also evaluate the BRL estimator described in Section 5.3.1. To obtain both, we estimate the posterior distribution of \mathbf{Z} using the Bayesian record linkage model of Sadinle (2017), which we summarize in Section 5.4.1. In Section 5.4.2, we consider a simulation study with varying quality of the variables used for matching and varying size of the overlap, i.e., the number of records in \mathcal{A} and \mathcal{B} where $\mathcal{A}_{i0} = \mathcal{B}_{j0}$. In Section 5.4.3, we consider genuine record linkage applications for which ground truth is available.

We investigate three questions. First, does using the F -score estimator result in higher quality record linkage performance? We evaluate this by computing the F -scores for the point estimates using ground truth data. Second, does using the F -score estimator provide accurate overlap size estimation? We evaluate this by computing the overlap size induced from the linkage estimate and comparing it to true overlap. Third, is using the F -score estimator compatible with model predictions, particularly credible intervals for the population size obtained from the Bayesian record linkage model? We evaluate this by assessing whether the estimated overlap is inside the 95% credible interval.

5.4.1 Bayesian Bipartite Record Linkage Model (♠)

Suppose each record $\mathcal{A}_i \in \mathcal{A}$ and $\mathcal{B}_j \in \mathcal{B}$ has F shared attributes, which we call linking fields. For any i and j , let $\mathcal{A}_i = (\mathcal{A}_{i,1}, \dots, \mathcal{A}_{i,F})$ and $\mathcal{B}_j = (\mathcal{B}_{j,1}, \dots, \mathcal{B}_{j,F})$, and let $\gamma_{i,j} = (\gamma_{i,j}^{(1)}, \dots, \gamma_{i,j}^{(F)})$ be the comparison vector derived from the linking fields for the record pair. Each $\gamma_{i,j}^{(f)}$ is the result of the comparison between $\mathcal{A}_{i,f}$ and $\mathcal{B}_{j,f}$. For instance, $\gamma_{i,j}^{(f)}$ can be a binary indicator of whether field f is identical for \mathcal{A}_i and \mathcal{B}_j . We assume that each $\gamma_{i,j}^{(f)}$ takes on $d_f \geq 2$ possible agreement levels.

We suppose that each $\gamma_{i,j}^{(f)}$ is a realization of a random variable $\Gamma_{i,j}^{(f)}$. Let $\Gamma_{i,j} = (\Gamma_{i,j}^{(1)}, \dots, \Gamma_{i,j}^{(F)})$. For $l = 1, \dots, d_f$, let $m_{f,l} = P(\Gamma_{ij}^{(f)} = l | Z_j = i)$ and $u_{f,l} = P(\Gamma_{ij}^{(f)} = l | Z_j \neq i)$ denote the probability that field f takes on value l for matches and non-matches, respectively. The model of Sadinle (2017) assumes the $\Gamma_{i,j}^{(f)}$ are conditionally independent given Z_j , so that

$$P(\Gamma_{i,j} = \gamma_{i,j} \mid Z_j = i) = \prod_{f=1}^F \prod_{l=1}^{d_f} m_{f,l}^{\mathbb{I}(\gamma_{i,j}^{(f)}=l)} \quad (5.26)$$

$$P(\Gamma_{ij} = \gamma_{ij} \mid Z_j \neq i) = \prod_{f=1}^F \prod_{l=1}^{d_f} u_{f,l}^{\mathbb{I}(\gamma_{i,j}^{(f)}=l)}. \quad (5.27)$$

The model specification is completed with (uniform) Dirichlet prior distributions for each $\mathbf{m}_f = (m_{f,1}, \dots, m_{f,d_f})$ and $\mathbf{u}_f = (u_{f,1}, \dots, u_{f,d_f})$, and a prior distribution on \mathbf{Z} that enforces the bipartite matching constraint Sadinle (2017).

The model parameters can be estimated using a Markov chain Monte Carlo sampler. This results in L plausible draws of the linkage structure as represented by $(\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(L)})$. In the simulations, we use the ‘‘BRL’’ package (Sadinle, 2020) available in the software R to estimate the posterior distribution and obtain these draws.

Table 5.3: The \mathbf{m} and \mathbf{u} parameters for the three simulation scenarios. Here, we present the probabilities that the fields match for linked records ($m_{f,1}$) and do not match for non-linked records ($u_{f,2}$). (♠)

Error Level	$m_{1,1}$	$m_{2,1}$	$m_{3,1}$	$u_{1,2}$	$u_{2,2}$	$u_{3,2}$
Low	0.93	0.93	0.98	0.94	0.94	0.98
Moderate	0.83	0.83	0.98	0.84	0.84	0.98
Moderate-High	0.83	0.83	0.88	0.84	0.84	0.98

5.4.2 Simulation Study (♠)

We consider several data scenarios characterized by different overlap sizes and error levels. To facilitate repeated sampling computations, we let $n_{\mathcal{A}} = 1000$ and $n_{\mathcal{B}} = 50$. We determine the overlap size by setting the proportion π of records in \mathcal{B} that have links in \mathcal{A} as $\pi \in \{25\%, 50\%, 75\%, 100\%\}$. In each scenario, we generate comparison vectors for $F = 3$ binary fields, where $\gamma_{i,j}^{(f)} = 1$ when $\mathcal{A}_{i,f} = \mathcal{B}_{j,f}$ and $\gamma_{i,j}^{(f)} = 2$ otherwise. For each field f , we set $\mathbf{m}_f = (m_{f,1}, m_{f,2})$ and $\mathbf{u}_f = (u_{f,1}, u_{f,2})$ to represent one of three error levels, where smaller values in $m_{f,1}$ accompanied by smaller values in $u_{f,2}$ indicate increased error levels in the linking fields. The parameter settings for each level are displayed in Table 5.3. These parameter choices are guided by the values of \mathbf{m}_f and \mathbf{u}_f from the RLdata500 data (Sariyar and Borg, 2022) that we use in Section 5.4.3.1. Briefly, the low error scenario represents minimal errors across all fields. The moderate and moderate-high error settings represent situations where errors are introduced into two of the three fields to varying extents. When referring to simulation results from the F -score point estimator and BRL estimator, we use F-Algo and BRL, respectively.

Table 5.4 displays the simulation results. For each metric, the averages are computed across 1000 independent simulation runs. Within any setting, the Monte Carlo standard errors for the average differences of the F -scores and the overlap sizes under BRL and F-Algo are below 10^{-6} . At low error levels, the BRL estimator

Table 5.4: Average F -score, estimated file overlap, true overlap, and model-based overlap 95% credible interval for 1000 replicates in the simulation study. F -score and estimated overlap are computed from the BRL and F-Algo point estimates. (♠)

Error Level	Overlap	F -score		Est. Overlap		True Overlap	Overlap 95% CI
		BRL	F-Algo	BRL	F-Algo		
Low	25%	.84	.88	13	14	13	(14,50)
	50%	.89	.91	27	28	24	(27,49)
	75%	.88	.86	33	35	37	(36,50)
	100%	.95	.95	46	46	50	(46,50)
Moderate	25%	0	.44	0	23	13	(0,48)
	50%	0	.45	0	28	24	(2,49)
	75%	0	.52	0	32	37	(11,50)
	100%	.60	.59	29	41	50	(34,50)
Moderate-high	25%	0	.24	0	20	13	(0,48)
	50%	0	.39	0	30	27	(0,48)
	75%	0	.50	0	31	37	(8,50)
	100%	.48	.50	25	38	50	(27,50)

and the F-Algo estimator have comparable performances, both in the F -scores and overlap size estimates. In scenarios with moderate to moderate-high noise, the F-Algo estimator consistently outperforms the BRL estimator in the sense of higher F -scores and more accurate overlap size estimates. In these settings, the BRL estimator is overly conservative. In fact, it often declares all potential pairs as non-links, leading to F -scores of zero. Not surprisingly, the absolute performance of both estimators declines as overlap size decreases. With the lowest overlap and at least moderate errors, F-Algo tends to assign too many record pairs as links, whereas BRL tends to declare all record pairs as non-links.

5.4.3 Illustrative Examples

In this section, we evaluate the F-Algo estimator using data where the ground truth is established. We begin by validating its performance on the RLdata500 (Sariyar

and Borg, 2022). RLdata500 is frequently used in the record linkage literature due to its simplicity and low amount of error (Steorts, 2015). We then test its effectiveness using the Union Army data (Fogel et al., 2000).

5.4.3.1 Linkage with the RLdata500

RLdata500 is a synthetic dataset comprising 500 personal information records, 50 of which are noisy duplicates. At most two records refer to the same person. The linking fields include components of names and birth date.

We construct a bipartite version of these data as follows. First, we split non-duplicated records at random between \mathcal{A} and \mathcal{B} . Second, for duplicated records, we place the first record instance in \mathcal{A} and the second instance in \mathcal{B} . For string attributes (first and last name), we construct $\gamma_{i,j}^f$ using normalized Levenshtein distance thresholds at $(0, 0.25, 0.5, 1)$, which is the default in the “BRL” package (Sadinle, 2020). For numeric attributes (birth year, birth month, and birth day), we use binary $\gamma_{i,j}^f$ to indicate exact match on each field.

We consider four versions of the Bayesian record linkage model of Section 5.4.1. The four models use different attributes for linkage. Model A only uses birth year, birth month, and birth day. Model B only uses the last name and birth year. Model C uses first name, last name, and birth year. Model D uses first name, last name, birth year, birth month, and birth day. We fit the models using the “BRL” package with default hyperparameters, collecting 20,000 posterior samples after a burn-in of 5,000 iterations.

Table 5.5 displays the results for the RLdata500 illustration. For Model A, Model C, and Model D, the F -score and estimated overlap size for F-Algo and BRL are identical. For Model B, the F-Algo estimate has a marginally higher F -score than the BRL estimate. In all cases, the estimated overlap size is inside the 95% credible interval. These observations are line with results from the simulation study, where the

Table 5.5: F -score, estimated file overlap, true overlap, and model-based overlap 95% credible interval for the four models estimated using RLdata500. F -score and estimated overlap are computed from the BRL and F-Algo point linkage estimates.

	F -score		Est. Overlap		True Overlap	Overlap 95% CI
	BRL	F-Algo	BRL	F-Algo		
Model A	0.71	0.71	32	32	50	(40, 122)
Model B	0.76	0.78	47	52	50	(37, 109)
Model C	0.90	0.90	43	43	50	(43, 58)
Model D	0.98	0.98	51	51	50	(49, 54)

two estimators offer similar results in the absence of substantial errors in the linking fields. Directly maximizing the expected F -score is compatible with anticipated behavior in these benchmark data.

5.4.3.2 Linkage with the Union Army Data

The Union Army data comprise a longitudinal sample of Civil War veterans collected as part of the Early Indicators of Aging project (Fogel et al., 2000). Records of soldiers from 331 Union companies were collected and carefully linked to a data file comprising military service records—which we call the MSR file—as well as other sources. These records also were linked to the 1850, 1860, 1900, and 1910 censuses. The quality of the linkages in this project is considered very high, as the true matches were manually made by experts (Fogel et al., 2000). Thus, the Union Army data file can be used to test automated record linkage algorithms.

We consider re-linking soldiers from the MSR data to records from the 1900 census, which we call the CEN data file. For the linking fields, we use first name, last name, middle initial, and approximate birth year. This linkage problem is difficult for automated record linkage algorithms due to the presence of soldiers’ family members in the CEN data. Furthermore, not all soldiers from the MSR data have a match in the CEN data.

We use two types of blocking to reduce the number of comparisons: on birth place, or on last name initial. For each scheme, we estimate the Bayesian record linkage model separately in the blocks. The first blocking scheme uses birth place. Here, we present results for a block comprising all records with birth place of Michigan; we call this Block 1. This block has 529 records in the MSR data and 1840 records in the CEN data. The second blocking scheme uses the first letter of the last name. Here, we present results for a block comprising all records with last name starting with “O” as in Osborn or O’Connell; we call this Block 2. This block has 504 records in the MSR data and 599 records in the CEN data. We note that blocking based on last name initial rather than birth place tends to result in more erroneous linkages. Individuals are more likely to have the same or a similar last name if they are grouped by last name initial than if they are grouped by birth place.

We consider two versions of comparison vectors, which results in two record linkage models. Model E uses Levenshtein distances to compare names with thresholds $(0, 0.25, 0.5, 1)$, a binary comparison for middle initial, and a three-level comparison with threshold for birth year. Here, we quantize differences in birth years to the bins $[0, 1]$, $(1, 5]$, and $(5, \infty)$. Model F is a slight modification, using the thresholds $(0, 0.1, 0.5, 1)$ for Levenshtein comparisons instead.

For each comparison vector set, we fit the Bayesian record linkage model described in Section 5.4.1 using the “BRL” package with default hyperparameters, collecting 20,000 posterior samples after a burn-in period of 5,000 iterations.

Table 5.6 displays the results from the linkage with the Union Army data. In the case of Block 1, the BRL and F-Algo estimates perform similarly, with the same F -score for Model E and only a marginal difference for Model F. For the estimated overlap, the F-Algo estimates are closer to the truth. In the case of Block 2, when there is more ambiguity, the F-Algo estimates have substantially higher F -scores than the BRL estimates. Furthermore, the estimated overlaps using F-Algo are inside

Table 5.6: F -score, estimated file overlap, true overlap, and model-based overlap 95% credible interval for the combinations of two models and two blocks for the Union Army data. F -score and estimated overlap are computed from the BRL and F-Algo point linkage estimates.

		F -score		Est. Overlap		True Overlap	Overlap
		BRL	F-Algo	BRL	F-Algo		95% CI
Block 1	Model E	0.87	0.87	154	161	188	(150, 178)
	Model F	0.74	0.75	117	138	188	(125, 161)
Block 2	Model E	0.51	0.66	58	88	144	(77, 110)
	Model F	0.23	0.57	22	88	144	(59, 99)

the 95% credible intervals, whereas the estimated overlaps using BRL are not. As seen previously, the F-Algo estimator tends to be less conservative than the BRL estimator in the presence of higher uncertainty, and thus can estimate the number of overlap links more accurately.

5.5 Discussion

We propose a post-processing algorithm for point estimation of the linkage structure in bipartite record linkage tasks. Given either a posterior distribution or pairwise match probabilities, the algorithm obtains a point estimate through approximately maximizing the expected F -score. The proposed optimization algorithm extends the approach in Jansche (2007) to bipartite record linkage. By exploiting the sparsity of the linkage matrix, we implement several computational efficiency improvements, ensuring that the algorithm achieves a satisfactory complexity bound. Results from the simulation study and illustrative applications highlight the potential for improved performance over other estimators when linking fields are measured with error.

One area for future research involves incorporating an explicit constraint for overlap size accuracy into the algorithm. Currently, optimizing over the expected

F -score implicitly balances precision and recall. However, it would be possible to integrate this balance directly into the score function, penalizing deviations of the precision-recall ratio from unity. Implementing such a constraint may improve the algorithm's performance, particularly in scenarios with small overlap size, where our F -score optimization approach is currently less effective. Additionally, while we focus on comparing F -score optimization with the BRL estimator, it also would be valuable to assess how the plug-in approach of (5.18) compares to other methods, such as the approaches in Fellegi and Sunter (1969) or Jaro (1989). Lastly, improvements can be made to enhance computational efficiency. A promising approach is the use of Bayesian or Lipschitz optimization algorithms to speed up the outer optimization step in (5.7). By identifying and applying an appropriate Lipschitz bound, we can efficiently exclude values of $\hat{\mathbf{C}}_{\text{Bayes}}(k)$ from the search space that do not lead to global maximums. This strategy is especially valuable in problems with large $n_{\mathcal{B}}$, where the discrete search space for the optimal $\hat{\mathbf{C}}_{\text{Bayes}}(k)$ may make it difficult to run the algorithm presented here efficiently.

6. Improving the Validity and Practical Usefulness of AI/ML Evaluations Using an Estimands Framework

Commonly, AI or machine learning (ML) models are evaluated on benchmark datasets. This practice supports innovative methodological research, but benchmark performance can be poorly correlated with performance in real-world applications—a construct validity issue. To improve the validity and practical usefulness of evaluations, we propose using an estimands framework adapted from international clinical trials guidelines. This framework provides a systematic structure for inference and reporting in evaluations, emphasizing the importance of a well-defined estimation target. We illustrate our proposal on examples of commonly used evaluation methodologies—involving cross-validation, clustering evaluation, and LLM benchmarking—that can lead to incorrect rankings of competing models (rank reversals) with high probability, even when performance differences are large. We demonstrate how the estimands framework can help uncover underlying issues, their causes, and potential solutions. Ultimately, we believe this framework can improve the validity of evaluations through better-aligned inference, and help decision-makers and model users interpret reported results more effectively.

6.1 Introduction

Evaluating AI or machine learning (ML) models is critical at all stages of ML projects, influencing both development and deployment phases (Reich and Barai, 1999; Schelter et al., 2015). It facilitates comparisons among algorithms, guides feature selection and training, and allows for iterative refinements while ensuring robust performance in production settings.

Commonly, models are evaluated by measuring performance on benchmark datasets (Liao et al., 2021). The practice has many limitations despite being a key contributor to fast progress in the field (Dehghani et al., 2021). In many disciplines, benchmark performance metrics often do not generalize well to real-world capability (Liao et al., 2021; Wang et al., 2022). Ferrari Dacrema et al. (2019) and Hutson (2020) documented “phantom progress,” where inappropriate use of benchmark datasets and baseline methods leads to misleading performance estimates and an illusion of progress. Oakden-Rayner et al. (2020) showed how “hidden stratification,” where meaningful subgroups are not identified in benchmark datasets, can lead to hidden failure modes that performance metrics fail to represent. More broadly, Hutchinson et al. (2022) observed that the “idealized breadth of evaluation concerns” is not reflected in common benchmark-based evaluation practices. These types of issues are sometimes referred to as construct validity issues, i.e. a misalignment between theoretical goals and practical measurement or inferential methods (Sjøberg and Bergersen, 2022; Biderman et al., 2024)

To help address these issues, we propose adapting the estimands framework from international clinical trials guidelines (ICH, 2019; Phillips and Clark, 2021) to ML evaluation. The goal of the framework is to better align evaluation objectives with the design of evaluations (e.g., how to acquire data and what measurements to make) and the data analysis (e.g., how to summarize results and how to make inferences).

It achieves this by emphasizing the importance of having well-defined targets of estimation, the estimands, to enable aligned and efficient evaluations. Without well-defined estimands, evaluation stops at taking measurements and cannot make meaningful generalizations or inferences, or cannot clearly report results that a broad community of users.

“Incorrect choice of estimand and unclear definitions for estimands lead to problems in relation to trial design, conduct and analysis and introduce potential for inconsistencies in inference and decision making.” (ICH Steering Committee, 2014)

The estimands framework formalizes statistical best practices, providing key steps to accurately describe the estimation target (the estimand) and emphasizing the subtler considerations that contribute to a meaningful definition. It is quite simple and straightforward, but nonetheless an important reminder and standardized structure for key components that must be considered in practice. Figure 6.1 provides an overview of the framework adapted to ML evaluation, illustrating the components of an estimand and its relationship with an evaluation objective and data analysis. More details are given in Section 6.4.

To support our proposal, we consider three examples that demonstrate failures of commonly used evaluation methodologies and how the estimands framework reveals causes and solutions. The examples are related to a fundamental evaluation problem: the accurate ranking of ML models according to a chosen dimension of performance. We define a performance rank reversal as occurring when a model is wrongly deemed superior to another, despite the opposite being true (see Section 6.3.1).

Our examples show that rank reversals can occur using commonly used evaluation methodologies in simple applications, despite substantial performance differences between models. They use common practices in the literature but are simplified to demonstrate that problems can arise even in the simplest situations. The three examples are:

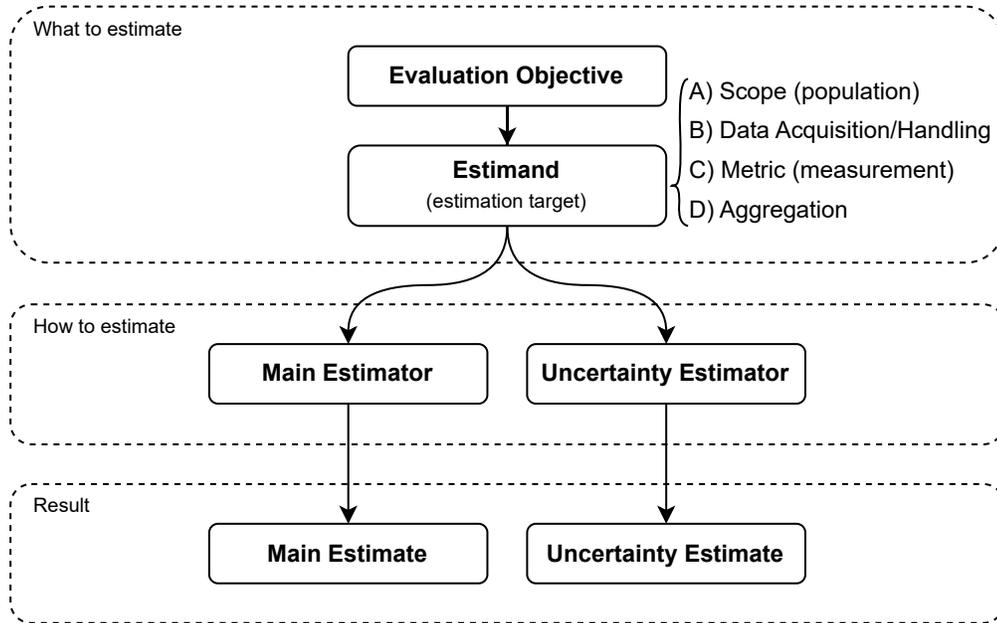


FIGURE 6.1: Estimands framework adapted from ICH (2019) for ML model evaluation, as described in Section 6.4. An evaluation objective is translated to an estimand. An estimand is characterized by (A) a metric or choice of measurement, (B) a specific scope (a population) to contextualize the metric, (C) a data acquisition strategy (including how missing data, data annotation inconsistencies, and other data issues are handled), and (D) an aggregation/summarization of the metric values over the given scope/population. Next, a main estimator is chosen to provide a sufficiently accurate estimate at minimal cost. Uncertainty regarding the estimation procedure can be separately or jointly estimated, accounting for sensitivity to the choice of the main estimator and its underlying assumptions.

Cross-Validation Example (Section 6.3.2): Unbiased cross-validation estimators (Stone, 1974; Bates et al., 2023) are widely used for model selection. We show in a simple regression example how cross-validation can lead to the selection of the worse model with high probability.

Clustering Evaluation Example (Section 6.3.3): Evaluating clustering models for entity resolution applications (Christophides et al., 2021), such as identity clustering based on face images (Shi et al., 2018), often relies on computing an F-score on a small benchmark dataset (Shi et al., 2018; Yin et al., 2020). We

show how the resulting F-score can be biased and unreliable for ranking models.

LLM Benchmarking Example (Section 6.3.4): We show how the composition of LLM benchmark datasets (Srivastava et al., 2022) along unmeasured dimensions can affect the relative performance of LLMs, making it difficult for rankings to generalize.

We apply the framework to each example to show its practical use in reviewing and developing evaluations. We also discuss some of the subtler issues involved in the definition of an estimand. Specifically, the application to the cross-validation example shows the importance of considering context and population for valid inferences. We use the clustering evaluation example to emphasize the impact of data acquisition issues on the definition of an estimand, and we use the LLM benchmarking example to discuss the potential of multi-criteria decision making methods

In summary, by using the estimands framework as scaffolding, we can ensure that ML evaluations are well aligned with key goals, that they produce valid inferences, and that their results are meaningful for applications and model users.

The rest of the paper is organized as follows. In Section 6.2, we provide background on ML evaluation and the approach of our paper by describing the importance of measurement, inference, and reporting in evaluations. Section 6.3 describes our three examples. We introduce our examples before the estimands framework to show how certain evaluation problems can be unexpected or surprising when they are not properly contextualized through the framework. Section 6.4 introduces the estimands framework and applies it to each example. Section 6.5 summarizes our findings and discusses broader potential for the estimands framework to improve ML evaluations.

6.2 Background

In Section 6.2.1, we define “models” and “evaluation,” and discuss common ML evaluation practices and their goals. In Section 6.2.2, we discuss our approach to ML evaluation.

6.2.1 Definitions and Related Work

We use model as an umbrella term for trained and untrained models, ML algorithms, and ML/AI systems. The scope is broad since we focus on statistical evaluation rather than any particular ML subfield. The statistical evaluation principles we discuss are widely used in applied statistics and other domains, such as clinical trials biostatistics. Therefore, we believe they are also useful for a wide range of ML applications.

We define evaluation as a study with the goal of making value judgments to guide action, decision, or change. See Wanzer (2021) for relevant discussion. We emphasize the scientific components of evaluation, and its goal of providing judgments that have practical consequences. The importance of judgemental evaluation is emphasized in Mathison (2005), and the importance of action-oriented outcomes is emphasized in Tong et al. (1987). For example, clinical trials aim to determine the efficacy, safety, and other characteristics of medical treatments, with a direct impact on clinical practice. Evaluation may be focused on developing cost-effective methods. For instance, adaptive designs are developed to reduce costs, improve accuracy, and improve patient outcomes in clinical trials.

There are three core components to evaluation: measurement, testing/inference, and reporting. Measurement captures a given characteristic of an object or state of the world. For example, in LLM evaluation, measurements are the scoring of responses on evaluation items. A large literature investigates techniques to efficiently and reliably score responses using human judges or automated methods (Liu et al.,

2023; Zhang et al., 2023; Zheng et al., 2024). Testing/inference involves checking assumptions and expectations, often probabilistically, to determine if we are likely right or wrong. In ML evaluation, inference might translate measurements from a training dataset to an expected generalization error. Statistical testing can assess the significance of performance differences between models (Dehghani et al., 2021). Reporting involves summarizing and communicating evaluation results, addressing the needs of its consumers. It bridges scientific insights and real-world change, requiring sufficient effort in summarization, communication, and analysis to support actions or change.

Too often, ML evaluations stop at measurement, only computing scores on a benchmark dataset (Post, 2018; Dehghani et al., 2021; Colombo et al., 2022; Srivastava et al., 2022). In these cases, there is often no uncertainty regarding the target of estimation and no direct consideration of how performance might generalize beyond the benchmark.

Even when inferences are made, say by estimating generalization performance through cross-validation, the process may not align with evaluation objectives. For instance, it is known that cross-validation estimators are often only weakly correlated, or even negatively correlated, with the generalization performance of a given model (Hastie et al., 2009; Bates et al., 2023). Whether or not a cross-validation estimate is representative of the generalization performance of a given trained model must be checked. This example is discussed in more detail in Section 6.4.2.1.

On the reporting front, standardized approaches like data cards and model cards are widely used (Mitchell et al., 2019; Pushkarna et al., 2022). But interpreting benchmarking results and performance evaluations can be difficult. Complex or unmeasured characteristics of a benchmark dataset impede understanding of what it represents and how results translate into practice (Dehghani et al., 2021). Another problem is the lack of reporting of disaggregated metrics or item-level performance

(Burnell et al., 2023). Furthermore, relatively little attention has been given to aggregating or reporting scores from multiple tasks (Colombo et al., 2022). Even with a clear evaluation objective, determining the adequacy of the evaluation methodology can be challenging. As shown in our examples (Section 6.3), many evaluation methodologies lead to rank reversals with high probability.

6.2.2 ML Evaluation as a Discipline

We treat ML evaluation as a dedicated discipline, focusing on applied performance estimation rather than model development. While evaluation is often tightly coupled with model development, it should also be considered as a separate topic to (1) ensure the robustness and efficiency of evaluations, (2) avoid conflicts of interest or misaligned incentives in high-risk applications, as emphasized in Board of Governors of the Federal Reserve System (2011), and (3) ensure that evaluation results are useful to decision-makers or model users. This separation is particularly important given the increased usage of general-purpose and pre-trained ML models across a wide range of software systems. In this context, evaluation is a separate activity closely tied to an application area rather than the development or refinement of a model. The large number of ML models used in many organizations motivates the development of efficient methodologies and systematic evaluation processes applicable across multiple domains.

6.3 Three Examples Leading to Rank Reversals

In this section, we define rank reversals and describe three examples where commonly used evaluation methods have high probability of rank reversal.

The examples provide motivation for the estimands framework, which is introduced next in Section 6.4. We delay explaining the cause for evaluation problems until Section 6.4, where we revisit each example to show how the framework provides the

necessary scaffolding to follow best practices and resolve issues. Our examples are intentionally simplified to highlight common evaluation problems, not best practices.

6.3.1 Defining Performance Rank Reversals

Consider two models, A and B , that we want to rank by a performance metric φ such as generalization accuracy. These models may be pre-trained and fixed, or viewed as random to account for variability in the training process or training data. We denote by $\varphi(A)$ and $\varphi(B)$ the scores of the two models under φ . A statistical evaluation methodology provides two corresponding estimators $\hat{\varphi}(A)$ and $\hat{\varphi}(B)$. Although estimating φ directly can be challenging in some situations, the ranking of alternatives generated by $\hat{\varphi}$ should reflect the ranking produced by φ , with sufficient probability. If not, the estimators are said to cause a rank reversal.

We formalize the general notion of rank reversals in Definition 7 below.

Definition 7 (Rank Reversals). Let φ be a performance metric of interest, and let A, B be two given models. The probability of rank reversal for the estimators $\hat{\varphi}(A)$ and $\hat{\varphi}(B)$ is defined as

$$\mathbb{P}\left(\text{rank}\{\hat{\varphi}(A), \hat{\varphi}(B)\} \neq \text{rank}\{\varphi(A), \varphi(B)\}\right), \quad (6.1)$$

where the probability taken with respect to the joint distribution of the estimators and the models A and B .

Note that the underlying probability space varies by context. In our cross-validation example of Section 6.3.2, we analyze properties of estimators from the perspective of hypothetical resampling of a training dataset and the following model training. In our clustering evaluation example of Section 6.3.3, we consider two fixed, pre-trained models, and randomness is introduced by labeling a sample of the data. In our LLM example of Section 6.3.4, everything is fixed and the probability of a rank reversal is either 0 or 1.

Rank reversals have significant consequences and costs in applications, beyond the direct impact of performance loss. For example, in the early stages of model development, rank reversals can lead to discarding useful features or modalities, thereby hindering future performance (Wang et al., 2022). These effects can compound throughout model development, leading to high costs.

6.3.2 Rank Reversals With Cross-Validation

6.3.2.1 Background on Cross-Validation

Consider a supervised learning problem with independent and identically distributed data points $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i \in \mathbb{N}$. A training algorithm \mathcal{A} maps a set of examples \mathcal{D} to a model represented by a function $\hat{f}_{\mathcal{D}} : \mathcal{X} \rightarrow \mathcal{Y}$. Given a loss function $\ell(\hat{y}, y) \geq 0$ and n examples $\mathcal{D}_n = \{(X_1, Y_i)\}_{i=1}^n$, the training algorithm aims to learn a function $\hat{f}_{\mathcal{D}_n}$ that minimizes the expected generalization error

$$\text{Err}_{\text{gen}}(\hat{f}_{\mathcal{D}_n}) = \mathbb{E} \left[\ell \left(\hat{f}_{\mathcal{D}_n}(X_{n+1}), Y_{n+1} \right) \mid \mathcal{D}_n \right]. \quad (6.2)$$

This generalization error cannot be computed exactly without knowledge of the distribution of (Y_i, X_i) . However, it is commonly estimated using cross-validation. For example, with leave-one-out cross-validation, we consider a point $(X, Y) \in \mathcal{D}_n$ sampled at random, a function $\hat{f}_{\mathcal{D}^*}$ learned from $\mathcal{D}^* = \mathcal{D}_n \setminus \{(X, Y)\}$, and we compute

$$\text{CV}_{\text{loo}}(\hat{f}_{\mathcal{D}_n}) = \mathbb{E} \left[\ell \left(\hat{f}_{\mathcal{D}^*}(X), Y \right) \mid \mathcal{D}_n \right] = \frac{1}{n} \sum_{i=1}^n \ell \left(\hat{f}_{\mathcal{D}_n \setminus \{(X_i, Y_i)\}}(X_i), Y_i \right). \quad (6.3)$$

The cross-validation estimator is considered “unbiased” in the sense that

$$\mathbb{E} \left[\text{Err}_{\text{gen}}(\hat{f}_{\mathcal{D}_n}) - \text{CV}_{\text{loo}}(\hat{f}_{\mathcal{D}_n}) \right] = 0. \quad (6.4)$$

Remark 11. It is well known that the cross-validation estimator $\text{CV}_{\text{loo}}(\hat{f}_{\mathcal{D}_n})$ is better at estimating the unconditional generalization error $\mathbb{E} \left[\text{Err}_{\text{gen}}(\hat{f}_{\mathcal{D}_n}) \right]$ than the

conditional generalization error $\text{Err}_{\text{gen}}(\hat{f}_{\mathcal{D}_n})$ (Hastie et al., 2009; Bates et al., 2023). However, in practice, the conditional generalization error is the target metric of interest since we care about the performance of a realized model. The key observations of Section 6.3.2.2 below do not change based on the choice of conditional or unconditional generalization error as a target metric. Rank reversals occur with high probability for both choices of target metrics.

6.3.2.2 Example: The California Housing Dataset

Unfortunately, cross-validation estimators can lead to a high probability of rank reversal, even in simple examples with large performance differences between two models.

We illustrate this problem in a simple, standard example. Consider the California Housing Dataset, which contains information on house attributes in $N = 20,640$ California block groups from the 1990 U.S. Census. For a given Census block group i , the response variable Y_i is the Census block group's median house value. The feature vector X_i has 8 numerical dimensions. The goal is to learn a predictive function for the median house value that minimizes the generalization performance, using a mean squared error loss function. We consider a training dataset \mathcal{D}_n of $n = 2,000$ examples, randomly selected with replacement from the set of California Census block groups, and two training algorithms: (A) a simple linear regression, and (B) a decision tree model with a maximum depth of 5, using default scikit-learn parameters for their implementation. These training algorithms were chosen for simplicity, rather than for accuracy or effectiveness. The true difference in performance between the two models is rather substantial. The root mean squared error of the linear model is around \$200,000, versus \$74,000 for the decision tree (the median house price ranges from \$30,000 to \$500,000).

Next, we compute the rank reversal probability for the leave-one-out cross-

Table 6.1: Summary of our cross-validation experiment. Decision trees are generally better at predicting median house price than linear models as seen by their lower average mean squared error (MSE). Furthermore, the average of the cross-validation (CV) estimates are close to the average MSE, showing that that CV estimates are nearly unbiased. However, in the majority of the replications (around 75% of them), the worst-performing model is selected.

	Linear Model	Decision Tree
Avg. of generalization MSE (φ)	4.02	0.553
Avg. of CV estimates	4.00	0.553
Probability of rank reversal	75.3%	

validation estimator CV_{loo} with respect to models resulting from sampling a training dataset \mathcal{D}_n and the following target performance metric:

φ : the expected generalization error Err_{gen} defined in (6.2), computed over all $N = 20,640$ California Census block groups.

We estimate the probability of rank reversal using 50,000 replications of sampling a training dataset of size $n = 2,000$, fitting each model, and comparing their leave-one-out cross-validation performance estimate to their true generalization performance.

The results of the experiment are summarized in Table 6.1, showing that the cross-validation estimator leads to a rank reversal with a probability of 75.3%.¹ Given this high probability of rank reversal, the cross-validation is not appropriate for choosing between linear and decision tree models in this application.

6.3.3 Rank Reversals in Clustering Evaluation

6.3.3.1 Background on Clustering Evaluation

Consider a clustering problem, where each element r of a set \mathcal{R} belongs to a cluster $c(r) \subset \mathcal{R}$, and where the resulting set of clusters \mathcal{C} partitions \mathcal{R} (each element r

¹ When choosing the unconditional generalization error $\mathbb{E}[\text{Err}_{\text{gen}}(\hat{f}_{\mathcal{D}_n})]$ as the target metric rather than the conditional generalization error, the probability of rank reversal is slightly lower at 66%.

belongs to a single cluster $c(r) \in \mathcal{C}$). The goal is to predict $\hat{c}(r) \in \mathcal{R}$ for the cluster to which each element r belongs, under the same constraint that the resulting set of predicted clusters $\hat{\mathcal{C}}$ partitions \mathcal{R} .

This can be equivalently formulated as a classification problem, where the goal is to predict whether two elements $r, r' \in \mathcal{R}$ belong to the same cluster under a transitivity constraint. Thus, classification evaluation metrics such as precision P , recall R , and the F-score F are commonly used to evaluate clustering models. Denoting by \mathcal{P} the set of pairs of elements predicted to belong to the same cluster and by \mathcal{T} the set of pairs that truly belong to the same cluster, these metrics are defined as

$$P = \frac{|\mathcal{T} \cap \mathcal{P}|}{|\mathcal{P}|}, \quad R = \frac{|\mathcal{T} \cap \mathcal{P}|}{|\mathcal{T}|}, \quad F = \left(\frac{P^{-1} + R^{-1}}{2} \right)^{-1}. \quad (6.5)$$

Since the true clustering is unknown, these metrics can only be computed for labeled benchmark datasets or labeled subsets of the data. A common practice in the literature is to compare clustering models based on the F-score computed on a benchmark dataset (Xie et al., 2013; Yin et al., 2020; Wang et al., 2022). In the context of our framework defined in Section 6.3.1, our target metric is $\varphi = F$, with the estimator $\hat{\varphi}$ being F-score computed on a benchmark dataset or a labeled data subset, and the clustering models considered to be pre-trained.

6.3.3.2 Example: Identity Clustering Based on Face Images

Unfortunately, the F-score tends to degrade with dataset size in a nonlinear manner, depending on specific characteristics of the clustering models. Thus, a clustering model A may outperform model B on small subsets of the data, but model B may outperform model A on the full dataset because it is less affected by performance degradation as dataset size increases (Binette et al., 2023). If performance on a labeled data subset is used to select a model for clustering the full data, this causes a

rank reversal: model A is chosen, even though model B performs better on the full dataset.

To illustrate this problem, we consider an identity clustering task related to facial recognition. Specifically, we use the Olivetti Faces dataset² which contains 400 face pictures of 40 individuals, with 10 images per person. The goal is to cluster the images by individual identity. To solve this task, we use k-means clustering on pre-trained FaceNet embeddings (Schroff et al., 2015; Esler, 2023). Model A uses $k = 30$ clusters and model B uses $k = 60$ clusters. The true F-score, which would be unknown in practice, can only be computed using the cluster membership of all 400 face pictures. It is estimated by computing the F -score on a benchmark dataset of 10 randomly selected individuals for whom the true clusters have been resolved. Table 6.2 shows the F -scores of the two models on the full dataset, the average F -score estimates, and the probability of rank reversal in 20,000 simulations of sampling 10 individuals to estimate the F-score. There is a 66% probability of rank reversal: the F-score estimator selects model B 66% of the time, even though model A performs better on the full dataset.

Table 6.2: Summary of the clustering evaluation experiment. Model A is better than model B in terms of F-score computed on the entire dataset (0.87 versus 0.73). However, the F-score estimator (computing the F-score on a random subset of 10 labeled clusters) is highly biased, leading to the worst model being selected around 66% of the time.

	Model A	Model B
True F-score	0.87	0.73
Avg. of F-score estimates	0.87	0.90
Probability of rank reversal	66%	

² The Olivetti Faces dataset was created at AT&T Lab Cambridge and obtained online from scikit-learn 1.3.

6.3.4 Rank Reversals in LLM Benchmarking

6.3.4.1 Background on Benchmarking Large Language Models

Methodological research on large language models (LLMs), like in other ML fields, relies on common benchmark datasets to evaluate models and track progress (Liao et al., 2021; Dehghani et al., 2021; Lin et al., 2021; Srivastava et al., 2022; Colombo et al., 2022; Zhou et al., 2023; Chang et al., 2023; Guo et al., 2023). Benchmark datasets are organized by task, topic, and other characteristics, with corresponding evaluation items. For example, the Format-Following benchmark dataset (Xia et al., 2024) tasks an LLM to follow formatting guidelines specified in a prompt. This benchmark dataset contains 494 evaluation items across 10 application domains, 50 subdomains, and 248 format types. Each evaluation item instructs the LLM to format data in a specified way. Success in format-following can be assessed by human annotators and/or a judging model.

The objective is to determine how well a given LLM should perform on tasks similar to the ones in the benchmark dataset. That is, the target metric φ is the generalization performance for a given task description, and the estimator $\hat{\varphi}$ is observed performance on the task’s benchmark.

There are two main challenges in evaluating LLMs. First, the open-endedness of expected answers. There is not always a single correct answer to a given evaluation item. In such cases, correctness is determined by a separate judge. Second, LLMs are evaluated for broad humanlike capabilities rather than for precisely defined quantitative objective.

In practice, these challenges are addressed by comparing LLMs to humans, effectively anthropomorphizing LLM capabilities (Chollet, 2019; Chang et al., 2023). LLMs are evaluated similarly to humans and are compared to humans to contextualize their performance (Srivastava et al., 2022; Chang et al., 2023). Consequently,

many implicit assumptions related to educational and psychological measurement are applied to LLMs to facilitate the interpretation of results. For example, statistical models used to quantify human performance typically assume a latent trait, such as ‘mathematical reasoning ability,’ estimated based on item responses. The scale of that latent variable may be ignored, as long as a well-defined population of reference can be used for z-scores. Through these latent traits, we can generalize to expected performance on similar tasks. For example, a human or LLM that performs better than others on a mathematical exam might be expected to perform better than others on similar mathematical tasks.

In this line of thinking, Hernández-Orallo (2017); Chollet (2019); Martínez-Plumed et al. (2019); Wang et al. (2023) proposed the use of psychometric methods for evaluating ML systems. For instance, item response theory (Cai et al., 2016; Lalor et al., 2024) is one way to formalize the estimation of latent traits from evaluation items. Note that there are limitations to traditional psychometric-based approaches, such as the reliance on a reference population, for which promising alternatives have been proposed (Hernández-Orallo et al., 2022; Burnell et al., 2022; Burden et al., 2023, 2024). In common practice, the use of LLM benchmarks is typically quite simple, only measuring the average performance on evaluation items. These performance averages are sometimes disaggregated by topic or by other characteristics.

6.3.4.2 Example: The Format-Following Benchmark Dataset

Unfortunately, current LLMs do not always behave like humans (see e.g., Efrat et al. (2022); Wang et al. (2023)), leading to unexpected performance rank reversals in some applications. LLM rankings for a given task’s benchmark dataset may not be maintained when considering other similar tasks, due to unmeasured confounders that would not be expected to impact human rankings to the same extent.

For instance, LLM rankings on the Format-Following benchmark are not always

Table 6.3: Example of a rank reversal between two models on the Format-Following benchmark dataset. We estimated question difficulty based on the average performance of a class of LLMs, leading to the following equally sized categories of questions: easy, hard, and expert. We ignore the expert category as the performance of open-source models is very low for these questions.

	Success Rate	
	Easy questions	Hard questions
Llama 2 7b	88%	36%
Mistral 7b instruct v0.1	80%	45%

stable across different difficulty levels (see table 6.3). An LLM may outperform another on difficult questions but perform worse on easy questions, or vice versa. Similar rank reversals can be observed in Mehrbakhsh et al. (2023) in the context of image recognition. As another example, Srivastava et al. (2022); Mizrahi et al. (2023) observed a class of LLMs being given different relative rankings based on a choice between semantically equivalent prompting templates. For a fluent English speaker, the prompt templates appear to be roughly equally straightforward, providing no reason to believe that performance ranks should differ based on template choice. In short, unknown characteristics of LLMs or benchmark datasets can affect their relative performance in unpredictable ways.

In our example, the scope of the “format-following” task is not well-defined. Even if it were, the unexpected sensitivity of rankings to certain characteristics of the benchmark dataset means that rank reversals are likely when these characteristics are not accounted for. Ideally, we would have a clear scope for the task and account for uncertainties associated with the selection of evaluation items, variability in evaluator scores, variability in model responses, and so forth.

6.4 Our Proposal: Better-Defined Targets of Estimation Through the Estimands Framework

We now describe the estimands framework adapted from ICH (2019) to improve the quality of inference and reporting in ML evaluations. The framework provides a structure for ML evaluations that emphasizes and clearly defines the target of estimation, i.e., the estimand. As we will see, precisely defining an estimand is a multi-step process that is more subtle than might appear at first glance. For instance, the counterfactual or potential outcome frameworks in causal inference are centered around defining the causal estimand (Höfler, 2005). ICH (2019) discusses many considerations that affect the treatment effect estimand in clinical trials, from the definition of the target population, to the handling of intercurrent events³. We discuss analogous issues in ML evaluation.

6.4.1 The Estimands Framework

We describe our proposed estimands framework as a series of steps to take in evaluations. Example applications are given next in Section 6.4.2.

6.4.1.1 Define the Evaluation Objective and Subject

The first step is to identify an evaluation objective and the subject of evaluation. The evaluation objective is the purpose of the evaluation, and the subject of evaluation can be one or more training algorithms, trained models, or machine learning systems.

6.4.1.2 Define the Estimand

The second step is to translate the evaluation objective into a precisely defined estimand. Defining an estimand requires a description of the following characteristics

³ Intercurrent events are “events during the study that may complicate the definition and estimation of the treatment effect estimand, such as premature discontinuation of randomized treatment, taking rescue medication, or death” (Darken et al., 2020)

(see Figure 6.1):

- (A) The scope or population of interest that the evaluation aims to cover, i.e. the context of application of an ML model. Clearly defining a scope or population of interest can be challenging. We discuss this further in Section 6.4.2.3 in application to LLM evaluation.
- (B) A data acquisition/handling strategy, describing how evaluation data is obtained and how data issues are handled. For example, if data annotators are employed, how are inconsistencies in labeling or missing labels handled? How reliable are the labels and are labeling errors addressed? How are dependencies between data points (e.g., temporal dependencies) accounted for? How representative is the evaluation data for the give population? Choices in the acquisition or handling of evaluation data impact what is being estimated and, consequently, how evaluation results should be interpreted and applied for decision-making. We discuss an example of the impact of these considerations in Section 6.4.2.2 in application to clustering evaluation.
- (C) A choice of metric (or metrics) to measure elements of the scope/population, such as squared error for regression problems.
- (D) A choice of aggregation method to summarize the behavior of the metric (A) over its scope (B), such as an arithmetic or geometric mean. When considering multiple metrics or measurements, the choice of aggregation procedure also concerns how they should be combined into an overall summary.

6.4.1.3 Choose Estimation Methodology

The third step is to choose an estimation methodology (in short, a main estimator) for the task at hand. As we show next in Section 6.4.2, the appropriateness of a given

estimation methodology depends on characteristics of the estimand and of evaluation data. Given evaluation data, the main estimator produces a main estimate, e.g., a numerical value that summarizes a model's performance.

6.4.1.4 Choose Uncertainty Estimation Methodology

The fourth step is to choose a method to estimate uncertainty (in short, an uncertainty estimator to provide a confidence interval) regarding the estimand. The goal is to account for uncertainty in the numerical value of the estimate that can be due to small sample sizes, variability in the evaluation data, labeling uncertainty, or uncertainty regarding assumptions that underlie the choice of evaluation methodology. In some cases, the main estimate and its uncertainty can be jointly estimated. In other cases, uncertainty may need to be characterized based on sensitivity analyses or other techniques. For example, one can account for sensitivity to estimation methods by reporting the range of estimates from different methodologies.

6.4.1.5 Reporting Standard

Given an evaluation that follows these steps, an external observer should be able to investigate whether or not the choice of estimand aligns with the evaluation objective, and whether the estimation methodology is adequate. They should also understand the scope of applicability of the evaluation's results and the extent to which results can transfer to other applications. For this, the components of the estimands framework should be put forward rather than relegated to footnotes and caveats.

6.4.2 Application of the Estimands Framework to Rank Reversal Examples

In this section, we walk through each of our rank reversal examples and apply our estimands framework from Section 6.4 to describe the target of estimation. Then, we show how the estimands framework helps explain the cause of the observed

rank reversals and suggest better or alternative evaluation approaches. We discuss incorporating data acquisition and handling issues in the definition of an estimand in Section 6.4.2.2.3. In Section 6.4.2.3.3, we discuss the use of multi-criteria decision-making methods (MCDM) to aggregate multiple scores into one and to help report on nuanced evaluation results.

Importantly, note that the contributions of this section are not the solutions and best practices that we mention. Rather, our goal is to show how the estimands framework can be used to systematically guide the practice and review of ML evaluations.

6.4.2.1 California Housing Dataset

Here are the components of the estimands framework given in the California Housing Dataset example of Section 6.3.2.2.

The evaluation objective is to estimate the expected generalization error of two models, with the ultimate goal of selecting the better model. We consider as subjects of evaluation a linear model and a decision tree model trained on a given dataset \mathcal{D}_n , $n = 2,000$.

The estimand is defined as follows. Our target population is all California Census block groups. The evaluation data is the training dataset, and other data acquisition/handling issues are ignored. The metric is the squared error, and we aggregate over the population using the mean.

As an estimator, we consider a leave-one-out cross-validation with no uncertainty quantification.

6.4.2.1.1 Cause of the Rank Reversals

We can identify the cause for the high probability of rank reversal by walking through the characteristics of the estimand and the models.

Regarding the target population, we have access to features for all California Census block groups. This can be used to verify that the evaluation data is representative of all block groups, or assess problems that can arise in predictions related to unrepresented subgroups. Figure 6.2 compares the distribution of features in an evaluation dataset to the full California dataset. We can see that many features have heavy-tailed distributions, and that block groups with outlying average occupation are not represented in the evaluation data. This is a problem given that the squared error metric and mean aggregation are both sensitive to extreme values.

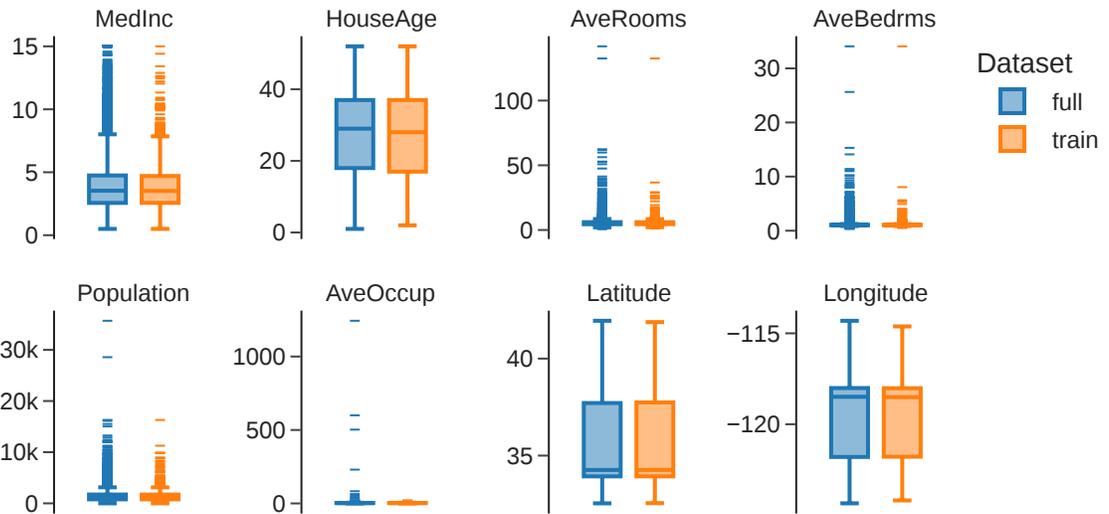


FIGURE 6.2: Comparing the distribution of features between the full California Housing Dataset and a training dataset of 2,000 random examples. Notice the four heavy-tailed features: the average number of rooms, the average number of bedrooms, the block group population, and the average occupation. Census block groups with outlying average occupation numbers are not represented in the training dataset.

Thinking about the models under consideration, we cannot expect a linear model to generalize well to outlying out-of-sample block groups, unless the linearity assumption was thoroughly justified and checked. In fact, in our example, we have median house price predictions in the negative millions for some block groups that are not

in the training data. These extremely inaccurate predictions are not reflected in cross-validation estimates of the linear model, leading to a high probability of rank reversal.

In short, we see that the cross-validation estimator is not suitable for the given estimand, given unrepresented outlying districts from the target population, the squared error metric, and the mean aggregation.

6.4.2.1.2 Potential Solutions to the Rank Reversals

There are multiple ways to address the rank reversal problem that we can identify by walking through the steps of the estimands framework and making appropriate changes.

First, we can question whether the choice of estimand is appropriate for the evaluation objective. This is highly application-specific but, given the heavy-tailed distribution of some block group features, we may prefer an estimand that provides more detailed insight into the behavior of models. For instance, we could separate city block groups from suburban and rural block groups, and aim to separately estimate generalization performance for each of these subgroups. If appropriate for a given application, alternative metrics and aggregations can be selected.

Second, assuming we keep our original estimand, we can question whether the training dataset was suitably selected to represent our target population. Given the existence of block groups with outlying features, a weighted sampling or stratified sampling scheme may be preferable to random sampling.

Third, we can rethink our estimation methodology. Given our model selection goal, we should complement any performance estimator with additional considerations that help inform an appropriate model selection decision. We can question the validity of assumptions that underlie our models, such as the linearity assumption. We can also check the validity of predictions on outlying block groups that are not in the training

data. Specifically, we know that median house prices are positive and we should be able to provide a plausible upper-bound based on subject-matter expertise. This can be used to check whether a model is generalizing properly to the full population of Census block groups. Alternatives to the cross-validation estimator may also be considered, such as confidence-based performance estimation (Białek et al., 2024), direct loss estimation (NannyML, 2024), or performance estimation through assessor models (Hernández-Orallo et al., 2022). These methods rely on features of the entire population, not just on features of the training dataset.

Finally, we may want to quantify uncertainty regarding cross-validation estimates in order to better understand the informativeness of these estimates for model selection. Unfortunately, methods that only rely on a training dataset (such as the cross-validation confidence interval method of Bates et al. (2023)) are not suitable here. Information about block groups with outlying features needs to be accounted for in order to detect the extrapolation problem associated with the linear regression model. Sensitivity analyses and simulation studies would be useful to characterize the accuracy of the cross-validation estimators in this example.

6.4.2.2 Clustering Evaluation

Here are the estimands framework components for the clustering evaluation example of Section 6.3.3.

The evaluation objective is to determine which of two models performs better to cluster a set of 400 face images that belong to 40 individuals. The metric is a binary indicator of whether a pair of two images represent the same individual or not, i.e., whether they match or not. That is, for two images i and j , we write $c_{i,j} = 1$ when i and j are a match and $c_{i,j} = 0$ otherwise, and we write $\hat{c}_{i,j}$ for the model prediction. This is then aggregated into an F-score, the harmonic mean of precision and recall. The scope or population of interest is the set of all $\binom{400}{2}$ image pairs (i, j) . Formally,

the F-score that we want to estimate can be written as

$$F = \frac{\sum_{i,j} c_{i,j} \hat{c}_{i,j}}{\left(\sum_{i,j} c_{i,j} + \sum_{i,j} \hat{c}_{i,j}\right)/2}. \quad (6.6)$$

The data acquisition strategy is to sample 10 clusters at random and to obtain the matching status of all pairs among them. We discuss practical data acquisition issues in more detail in a subsection below.

6.4.2.2.1 Cause for Rank Reversals

Going through the components of the estimands framework, the cause for rank reversals quickly becomes apparent.

Considering the target population (the full dataset), when sampling 10 clusters at random, the ratio of matching to non-matching pairs in the evaluation data is not representative of the full dataset. Indeed, in the full dataset of 400 images, there are $\binom{10}{2} = 45$ matching pairs for each of the 40 individuals, and 78,000 non-matching pairs in total, resulting in a match to non-match ratio of $1,800/78,000 \approx 2.3\%$. In a sample of 10 random clusters, there are 450 matching pairs, and 4500 non-matching pairs, resulting in a matching ratio of 10%.

The difference in the distribution of classes creates a bias in precision estimates, but not in recall estimates. As such, rank reversals in for the aggregate F-score can occur when comparing models that do not have the same recall. The problem is studied extensively in Foxcroft et al. (2024).

6.4.2.2.2 Solution to the Rank Reversals

A solution from Binette et al. (2023, 2024) is to change how the estimand is described in order to facilitate its estimation from a cluster sample. To do so, the F-score is expressed as a function of cluster metrics rather than pairwise errors, and this representation is then used to derive an estimator that is unbiased for random

cluster sampling. That is, we consider the population of clusters, rather than the population of pairs of images, and we change the metric and aggregation to still obtain the same F-score as a result. A random sample of clusters is then representative of the population and accurate estimates can be obtained from the estimators described in Binette et al. (2024).

6.4.2.2.3 Practical Data Acquisition and Handling Issues

In practice, we often need data annotators to label data, here to identify “ground truth” clusters. This is a difficult task when true cluster membership cannot be verified exactly. As such, biases, inaccuracies, and inconsistencies in the data labeling, and how they are handled, affect what is being estimated.

For example, in Binette et al. (2023), we evaluated a large-scale identity clustering system for patent inventors based on a probabilistic random sample of 400 ground truth clusters identified by a team of data annotators. In order to facilitate the process, data annotators could access the system’s current clustering predictions and use them as a starting point of the labeling, obtaining true clusters by cleaning and merging predicted clusters as needed. This approach created a bias towards current predictions, since they were used as a default.

In other words, we were not estimating the “true” F-score of the system. We were estimating the difference (summarized by a F-score) between current predictions and a manually corrected version of these predictions, where only sufficiently obvious errors were accounted for. This estimand that should have been directly stated, instead of mentioning biases as a potential caveat of the estimation methodology. This estimand that accounts for data acquisition issues is clear, unambiguous, and still perfectly relevant to the objective of improving predictions.

In short, we believe that an estimand needs to account for data acquisition and handling issues in order to be well-defined. Even though we would ideally want to

estimate φ , if data acquisition and handling issues cause us to estimate φ' , then the latter needs to be stated as the estimand. This way, the results of an evaluation can be properly interpreted without having to analyze in detail the estimation methodology, its assumptions, and its caveats.

6.4.2.3 LLM Evaluation

Here are the components of the estimands framework given in the example of Section 6.3.4. Our goal is to determine which of two LLMs is better at following a wide range of formatting instructions. Our metric is a binary indicator of success on evaluation items, and this is aggregated into an unweighted average success rate. The scope of formatting instructions is implicitly defined through the Format-Following benchmark dataset. Data acquisition and handling issues are ignored.

6.4.2.3.1 Cause of Rank Reversals

The estimand is too vaguely defined for the evaluation objective, since it depends on important but unmeasured characteristics of the benchmark dataset. Relatively small variations in the composition of the benchmark dataset can lead to unexpected rank reversals, due to the fact that the mean aggregation does not account for characteristics of the data. Without doing an in-depth analysis of the benchmark dataset, we don't know what the rankings represent, and we don't know to what kind of format-following tasks they would generalize.

6.4.2.3.2 What Is a Solution?

A solution is to better define the estimand through the specification of a clear scope for the format-following task. For instance, a probability distribution of format-following prompts can be specified through prompt templates or a generative prompt model. This can clarify the composition of the benchmark dataset at a high level and

help understand the applicability of evaluation results. This aligns with the following recommendation from Davis (2023):

“Benchmark sets should be constructed using a well-defined and replicable methodology. If one considers the process of running some AI system S on benchmark B to be an experiment testing a hypothesis, then the hypothesis ‘System S achieves performance P over problems with characteristics $X, Y, Z,$ ’ is a considerably more meaningful and cogent statement than ‘System S achieves performance P on the specific benchmark $B.$ ’ If B can be claimed to be a representative or a random selection of problems with characteristics S, Y, Z then there is some support for the stronger statement. If there is only the benchmark set B constructed catch-as-catch-can, then it is hard to know how these results will generalize. In practice, this is rare, except for benchmarks created using automatic synthesis.”

Now, it could be difficult to define a single scope that is representative a broad range of applications. Instead, we can evaluate on multiple narrow scopes, and then aggregate the results through multi-criteria decision-making (MCDM) methods (see next section). For example, we can separately evaluate performance for different format types (e.g. JSON, markdown), for different prompting templates, at different difficulty levels, and so forth. If a single model performs best across all tasks, it can be declared best. Otherwise, we can use MCDM methods to help provide a single score aggregate and detailed evaluation results that are useful for a variety of applications.

More refined approaches to the definition of an estimand might use psychometric methods. For instance, Burden et al. (2024) introduces “measurement layouts,” a framework to characterize system performance through the relationship between system capabilities and the characteristics of evaluation items (Burden et al., 2024). Key in this framework is the estimation of clearly scoped capabilities through the use of evaluation item characteristics. The estimated capability characteristics of a system can then be used to predict performance on new items.

6.4.2.3.3 Multi-Criteria Decision-Making and the Pareto Frontier

Multi-criteria decision-making (MCDM) (Triantaphyllou et al., 1998; Greco et al., 2016) is the task of identifying an optimal solution to a problem given multiple criteria that can conflict with one another. For example, we may want to choose an LLM that balances performance across many different tasks. Often, no single solution is better than alternatives for all criteria, resulting in a set of non-dominated solutions called the Pareto frontier (see Martínez-Plumed et al. (2018); Raji and Buolamwini (2024) for the use the Pareto Frontier in AI evaluation).

The Pareto frontier can be easily visualized when two or three criteria are considered. When more criteria are considered, dimensionality reduction and data visualization techniques can be used to summarize the behavior of alternatives across criteria (Ibrahim et al., 2016).

Another focus of MCDM is the translation of subjective preferences into an acceptable decision. For instance, the Analytic Hierarchy Process (AHP) (Saaty, 2003) translates pairwise relative preferences between criteria into a weighting scheme for computing a weighted average. There is never a single “true” or “best” solution to a choice of score aggregation. This is always a subjective process. The key is for the process to be transparent and for useful information to be provided.

6.5 Discussion

We highlighted current issues in the evaluation of ML models: (1) validity issues due to a lack of consideration of inference and reporting as key components of evaluation, beyond merely computing metrics, and (2) a high probability of rank reversals in specific cross-validation, clustering evaluation, and LLM evaluation applications. To address these problems, we proposed emphasizing the role of estimands (the targets of estimation) in ML evaluation, applying the estimands framework from

clinical trials biostatistics. We showcased how defining estimands through four key components helps identify methodological problems, highlight appropriate estimation methodologies, and properly interpret evaluation results.

Using an estimands framework to clearly define targets of estimation and to structure evaluations unlocks a number of key benefits. First, estimands enable the use of more sophisticated statistical methodologies to reduce the cost of evaluations. For instance, scoring LLMs on a large set of questions can be expensive. Through an estimand that specifies a clear population of questions, we can efficiently sample a subset that provides a sufficiently accurate performance estimate at a lower cost. Second, well-defined estimands enable uncertainty quantification. Without an estimand, we are simply computing metrics for which there is no context for uncertainty. An estimand provides a meaningful target of inference regarding which probability statements can be made. Finally, requirements for the definition of an estimand and for the structure of evaluations can be used in AI governance and AI auditing to help verify the quality and relevance of evaluations.

7. Conclusion

This dissertation considered statistical aspects of two important tasks: data linkage and model (or machine learning system) evaluation. Addressing methodological problems in these areas contributes to better data and better models for analytical systems and decision-making.

Bibliography

- Aggarwal, A., P. Lohia, S. Nagar, K. Dey, and D. Saha (2019). Black box fairness testing of machine learning models. *ESEC/FSE 2019 - Proceedings of the 2019 27th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 625–635.
- Aleshin-Guendel, S. (2020). On the identifiability of latent class models for multiple-systems estimation. *arXiv preprint arXiv:2008.09865*.
- Aleshin-Guendel, S., M. Sadinle, and J. Wakefield (2021). Revisiting identifying assumptions for population size estimation. *arXiv preprint arXiv:2101.09304*.
- Allain, J. (2017). Contemporary slavery and its definition in law. In A. Bunting and J. Quirk (Eds.), *Contemporary Slavery*, pp. 36–66. University of British Columbia Press.
- Amstrup, S. C., T. L. McDonald, and B. F. J. Manly (2005). *Handbook of Capture-Recapture Analysis*. Princeton, N.J.: Princeton University Press.
- Bagga, A. and B. Baldwin (1998). Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation*, Volume 1, pp. 563–566.
- Bai, E. A., O. Binette, and J. P. Reiter (2023). Optimal F-score clustering for bipartite record linkage. *arXiv preprint arXiv:2311.13923*.
- Bailey, M. J., C. Cole, M. A. C. Henderson, and C. G. Massey (2017). *How Well Do Automated Methods Perform in Historical Samples?: Evidence from New Ground Truth*. National Bureau of Economic Research.
- Baillargeon, S., L.-P. Rivest, et al. (2007). Rcapture: Loglinear models for capture-recapture in R. *Journal of Statistical Software* 19(5), 1–31.
- Bales, K., O. Hesketh, and B. W. Silverman (2015). Modern slavery in the UK: How many victims? *Significance* 12(3), 16–21.
- Bales, K., L. T. Murphy, and B. W. Silverman (2019). How many trafficked people are there in Greater New Orleans? lessons in measurement. *Journal of Human Trafficking* 6, 375–387.

- Balsmeier, B., M. Assaf, T. Chesebro, G. Fierro, K. Johnson, S. Johnson, G.-C. Li, S. Lück, D. O’Reagan, B. Yeh, et al. (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.
- Balsmeier, B., A. Chavosh, G.-C. Li, G. Fierro, K. Johnson, A. Kaulagi, D. O’Reagan, B. Yeh, and L. Fleming (2015). Automated disambiguation of us patent grants and applications. *Unpublished Working Paper*.
- Barnes, M. (2015). A practitioner’s guide to evaluating entity resolution results. *arXiv preprint arXiv:1509.04238*.
- Bates, D., M. Maechler, and M. Jagan (2023). Matrix: Sparse and dense matrix classes and methods. <https://CRAN.R-project.org/package=Matrix>.
- Bates, S., T. Hastie, and R. Tibshirani (2023). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 1–12.
- Belin, T. R. and D. B. Rubin (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association* 90(430), 694–707.
- Betancourt, B., J. Sosa, and A. Rodríguez (2022). A prior for record linkage based on allelic partitions. *Computational Statistics & Data Analysis* 172, 107474.
- Białek, J., W. Kuberski, and N. Perrakis (2024). We don’t need no labels: Estimating post-deployment model performance under covariate shift without ground truth. *arXiv preprint arXiv:2401.08348*.
- Biderman, S., H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, et al. (2024). Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.
- Bilenko, M. and R. J. Mooney (2003). On evaluation and training-set construction for duplicate detection. *Proceedings of the KDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 7–12.
- Binette, O. (2022). ER-Evaluation: An end-to-end evaluation framework for entity resolution systems. <https://github.com/OlivierBinette/ER-Evaluation>.
- Binette, O., Y. Baek, S. Engineer, C. Jones, A. Dasyuva, and J. P. Reiter (2024). How to evaluate entity resolution systems: An entity-centric framework with application to inventor name disambiguation. *arXiv preprint arXiv:2404.05622*.
- Binette, O. and J. P. Reiter (2023). ER-Evaluation: End-to-End Evaluation of Entity Resolution Systems. *Journal of Open Source Software* 8(91), 5619.

- Binette, O. and R. C. Steorts (2022a). (Almost) all of entity resolution. *Science Advances* 8(12), eabi8021.
- Binette, O. and R. C. Steorts (2022b). On the reliability of multiple systems estimation for the quantification of modern slavery. *Journal of the Royal Statistical Society Series A: Statistics in Society* 185(2), 640–676.
- Binette, O., S. A. York, E. Hickerson, Y. Baek, S. Madhavan, and C. Jones (2023). Estimating the performance of entity resolution algorithms: Lessons learned through patentsview.org. *The American Statistician* 77(4), 370–380.
- Bird, S. M. and R. King (2018). Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Annual Review of Statistics and Its Application* 5(1), 95–118.
- Bishop, Y. M., S. E. Fienberg, and P. W. Holland (2007). *Discrete Multivariate Analysis: Theory and Practice*. New York, NY: Springer.
- Board of Governors of the Federal Reserve System (2011). Supervision and regulation letter 11-7. Guidance on Model Risk Management.
- Bohning, D., P. G. M. Van der Heijden, and J. Bunge (2017). *Capture-Recapture Methods for the Social and Medical Sciences*. Boca Raton: CRC Press.
- Braiek, H. B. and F. Khomh (2020). On testing machine learning programs. *Journal of Systems and Software* 164, 110542.
- Brittain, S. and D. Böhning (2009). Estimators in capture-recapture studies with two sources. *ASTA Advances in Statistical Analysis* 93(1), 23–47.
- Bunke, T. (2016). Human trafficking legislation as a resource: Contradictory interpretations of human trafficking in zambia. *Journal of Trafficking, Organized Crime and Security* 2(2), 113–126.
- Burden, J., L. Cheke, J. Hernández-Orallo, M. Tešić, and K. Voudouris (2024). Measurement layouts for capability-oriented AI evaluation. AAAI: Tutorial.
- Burden, J., K. Voudouris, R. Burnell, D. Rutar, L. Cheke, and J. Hernández-Orallo (2023). Inferring capabilities from task performance with Bayesian triangulation. *arXiv preprint arXiv:2309.11975*.
- Burnell, R., J. Burden, D. Rutar, K. Voudouris, L. Cheke, and J. Hernández-Orallo (2022). Not a number: Identifying instance features for capability-oriented evaluation. In *International Joint Conference on Artificial Intelligence*, pp. 2827–2835.

- Burnell, R., W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela, M. Shanahan, E. M. Voorhees, A. G. Cohn, J. Z. Leibo, and J. Hernandez-Orallo (2023). Rethink reporting of evaluation results in AI. *Science* 380(6641), 136–138.
- Böhning, D. (2020). Discussion of read paper “multiple-systems analysis for the quantification of modern slavery: Classical and Bayesian approaches”. *183*(3), 716.
- Cai, L., K. Choi, M. Hansen, and L. Harrell (2016). Item response theory. *Annual Review of Statistics and Its Application* 3, 297–321.
- Canada, S. (2022). Social data linkage environment. Accessed: November 27, 2022, <https://www.statcan.gc.ca/en/sdle/index>.
- Canada, S. (Ed.) (2023). *Agriculture–Population Linkage: Data Quality Report, 2021*. Catalogue no. 32260006. Statistics Canada.
- Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1), 1–32.
- Chan, L., B. W. Silverman, and K. Vincent (2020). Multiple systems estimation for sparse capture data: Inferential challenges when there are nonoverlapping lists. *Journal of the American Statistical Association*, 1–10.
- Chandola, V., A. Banerjee, and V. Kumar (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41(3), 1–58.
- Chang, Y., X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.
- Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 265–270.
- Chao, A., N. J. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84(1), 45–67.
- Chao, A., H. Y. Pan, and S. C. Chiang (2008). The Petersen-Lincoln estimator and its extension to estimate the size of a shared population. *Biometrical Journal* 50(6), 957–970.
- Chao, A., P. K. Tsay, S. H. Lin, W. Y. Shau, and D. Y. Chao (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* 20(20), 3123–3157.

- Chao, A., Y. T. Wang, and L. Jost (2013). Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution* 4(11), 1091–1100.
- Chollet, F. (2019). On the measure of intelligence. *arXiv Preprint arXiv:1911.01547*.
- Chorev, S., P. Tannor, D. B. Israel, N. Bressler, I. Gabbay, N. Hutnik, J. Liberman, M. Perlmutter, Y. Romanyshyn, and L. Rokach (2022). Deepchecks: A library for testing and validating machine learning models and data. *Journal of Machine Learning Research* 23(285), 1–6.
- Choudhury, P. and D. Y. Kim (2019). The ethnic migrant inventor effect: Codification and recombination of knowledge across borders. *Strategic Management Journal* 40(2), 203–229.
- Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag.
- Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review*.
- Christen, P. and P. Christen (2012). *The data matching process*. Springer.
- Christen, P. and K. Goiser (2007). Quality and complexity measures for data linkage and deduplication. *Studies in Computational Intelligence* 43, 127–151.
- Christophides, V., V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis (2021). An overview of end-to-end entity resolution for big data. *ACM Computing Surveys* 53(6), 1–2.
- Chuang, J. A. (2014). Exploitation creep and the unmaking of human trafficking law. *American Journal of International Law* 108(4), 609–649.
- Chung, Y., T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang (2019). Slice finder: Automated data slicing for model validation. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1550–1553.
- Cochran, W. G. (1977). *Sampling Techniques*. New York: John Wiley.
- Cockayne, J. (2015). *Unshackling Development: Why We Need a Global Partnership to End Modern Slavery*. United Nations University.
- Colombo, P., N. Noiry, E. Irurozki, and S. Cléménçon (2022). What are the best systems? new perspectives on NLP benchmarking. *Advances in Neural Information Processing Systems* 35, 26915–26932.
- Cormack, R. M. (1968). The statistics of capture-recapture methods. *Oceanography and Marine Biology: An Annual Review* 6(1), 55–506.

- Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* 45(2), 395–413.
- Cormack, R. M. (1999a). Problems with using capture-recapture in epidemiology: An example of a measles epidemic. *Journal of Clinical Epidemiology* 52(10), 909–914.
- Cormack, R. M. (1999b). Reply to preceding comments. *Journal of Clinical Epidemiology* 52(19), 929–933.
- Cormack, R. M. (2000). Response. *Journal of Clinical Epidemiology* 53, 1275–1277.
- Dalzell, N. M. and J. P. Reiter (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics* 27(4), 728–738.
- D’Amour, A., K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.
- Darken, P., J. Nyberg, S. Ballal, and D. Wright (2020). The attributable estimand: A new approach to account for intercurrent events. *Pharmaceutical Statistics* 19(5), 626–635.
- Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics* 8(3), 522–539.
- Dasylyva, A., M. Abeysundera, B. Akpoue, M. Haddou, and A. Saidi (2016). Measuring the quality of a probabilistic linkage through clerical reviews. In *Proceedings of the 2016 International Methodology Symposium*.
- Dasylyva, A., M. Abeysundera, M. Haddou, and M. Lachance (2015). Sampling and estimation based on a frame of connected record groups. Statistics Canada Internal Report.
- Dasylyva, A., S. Canada, and A. Goussanou (2020). Estimating the false negatives due to blocking in record linkage. (February), 0–15.
- Dasylyva, A. and A. Goussanou (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science* 5, 181–216.
- Datta, M. N. and K. Bales (2013). Slavery in europe: Part 1, estimating the dark figure. *Human Rights Quarterly* 35(4), 817–829.
- Davidson, J. O. (2015). *Modern Slavery: The Margins of Freedom*. New York, NY: Palgrave Macmillan.

- Davies, D. L. and D. W. Bouldin (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1* (2), 224–227.
- Davis, E. (2023). Benchmarks for automated commonsense reasoning: A survey. *ACM Computing Surveys* 56(4), 1–41.
- Dawid, A. P. and S. L. Lauritzen (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics* 21(3), 1272–1317.
- Dehghani, M., Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, and O. Vinyals (2021). The benchmark lottery. *arXiv Preprint arXiv:2107.07002*.
- DiCiccio, T. J. and B. Efron (1996). Bootstrap confidence intervals. *Statistical Science* 11(3), 189–228.
- Doherr, T. (2017). Inventor mobility index: A method to disambiguate inventor careers. *SSRN Electronic Journal* (17).
- Doherr, T. (2021a). Disambiguation by namesake risk assessment. *ZEW-Centre for European Economic Research Discussion Paper* (21-021).
- Doherr, T. (2021b). Disambiguation by namesake risk assessment. *SSRN Electronic Journal* (21).
- Dong, X. L. and D. Srivastava (2015). *Big Data Integration*. Morgan and Claypool Publishers.
- Dottridge, M. (2017). Eight reasons why we shouldn’t use the term ‘modern slavery’. Accessed February 2, 2020, <https://www.opendemocracy.net/en/beyond-trafficking-and-slavery/eight-reasons-why-we-shouldn-t-use-term-modern-slavery/>.
- Draisbach, U. and F. Naumann (2013). On choosing thresholds for duplicate detection. *Proceedings of the 18th International Conference on Information Quality, ICIQ 2013*.
- Duran, B. S. and P. L. Odell (2013). *Cluster Analysis: A Survey*, Volume 100. Springer Science & Business Media.
- Dyché, J. and E. Levy (2006). *Customer Data Integration: Reaching a Single Version of the Truth*. New York: John Wiley & Sons.
- Earl, D. J. and M. W. Deem (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* 7(23), 3910–3916.
- Efrat, A., O. Honovich, and O. Levy (2022). Lmentry: A language model benchmark of elementary language tasks. *arXiv Preprint arXiv:2211.02069*.

- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Elastic (2022). Elasticsearch. <https://github.com/elastic/elasticsearch>.
- Enamorado, T., B. Fifield, and K. Imai (2019a). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review* 113, 353–371.
- Enamorado, T., B. Fifield, and K. Imai (2019b). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review* 113(2), 353–371.
- Esler, T. (2023). facenet-pytorch. <https://github.com/timesler/facenet-pytorch>.
- Far, S. S., R. King, S. Bird, A. Overstall, H. Worthington, and N. Jewell (2021). Multiple systems estimation for modern slavery: Robustness of list omission and combination. *Crime & Delinquency* 67(13-14), 2213–2236.
- Farcomeni, A. and L. Tardella (2012). Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electronic Journal of Statistics* 6, 2602–2626.
- Farrel, A., M. Dank, M. Kafafian, S. Lockwood, R. Pfeffer, A. Hughes, and K. Vincent (2019). Capturing human trafficking victimization through crime reporting. *National Institute of Justice*, 1–39.
- Feingold, D. A. (2010). Trafficking in numbers: The social construction of human trafficking data. In P. Andreas and K. Greenhill, M. (Eds.), *Sex, Drugs, and Body Counts*, pp. 46–74. Ithaca: Cornell University Press.
- Fellegi, I. P. and A. B. Sunter (1969). A theory for record linkage. *Journal of the American Statistical Association* 64(328), 1183–1210.
- Ferrari Dacrema, M., P. Cremonesi, and D. Jannach (2019). Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109.
- Ferreira, A. A., M. A. Gonçalves, and A. H. F. Laender (2012). A brief survey of automatic methods for author name disambiguation. *ACM Sigmod Record* 41(2), 15–26.
- Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* 59(3), 591–603.

- Fogel, R. W., D. L. Costa, M. Haines, C. Lee, L. Nguyen, C. Pope, I. Rosenberg, N. Scrimshaw, J. Trussell, S. Wilson, et al. (2000). Aging of veterans of the Union Army: Version M-5. *Chicago: Center for Population Economics, University of Chicago Graduate School of Business, Department of Economics, Brigham Young University, and the National Bureau of Economic Research*.
- Fortini, M., B. Liseo, A. Nuccitelli, and M. Scanu (2001). On Bayesian record linkage. *Research in Official Statistics* 4(1), 185–198.
- Foxcroft, J., P. Christen, and L. Antonie (2024). Class ratio and its implications for reproducibility and performance in record linkage. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 194–205. Springer.
- Friedman, J., T. Hastie, R. Tibshirani, et al. (2001). *The Elements of Statistical Learning*, Volume 1. New York, NY: Springer Series in Statistics.
- Frisoli, K. and R. Nugent (2018). Exploring the effect of household structure in historical record linkage of early 1900s ireland census records. In *Proceedings of the 2018 IEEE International Conference on Data Mining Workshops*, pp. 502–509. IEEE.
- Fuller, W. A. (2011). *Sampling Statistics*. John Wiley.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. Boca Raton: CRC Press.
- Gelman, A., D. B. Rubin, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7(4), 457–472.
- Greco, S., J. Figueira, and M. Ehrgott (2016). *Multiple criteria decision analysis* (2 ed.). Springer.
- Gu, G., S. Lee, and J. Kim (2008). Matching accuracy of the Lee-Kim-Marschke computer matching program.
- Guha, S., J. P. Reiter, and A. Mercatanti (2022). Bayesian causal inference with bipartite record linkage. *Bayesian Analysis* 17, 1275 – 1299.
- Guo, Z., R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong, et al. (2023). Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.
- Gutman, R., C. C. Afendulis, and A. M. Zaslavsky (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association* 108, 34–47.

- Hald, A. (2005). *A History of Probability and Statistics and Their Applications Before 1750*. New York, NY: Wiley.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York, NY: Springer-Verlag.
- Han, H., Y. Yu, L. Wang, X. Zhai, Y. Ran, and J. Han (2019). Disambiguating USPTO inventor names with semantic fingerprinting and DBSCAN clustering. *The Electronic Library* 37(2), 225–239.
- Hand, D. and P. Christen (2018). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing* 28(3), 539–547.
- Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Volume 2. Springer.
- Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review* 48, 397–447.
- Hernández-Orallo, J., W. Schellaert, and F. Martínez-Plumed (2022). Training on the test set: Mapping the system-problem space in AI. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 36, pp. 12256–12261.
- Herzog, T., F. Scheuren, and W. Winkler (2007). *Data Quality and Record Linkage Techniques*. New York, NY: Springer.
- Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology* 5, 1–12.
- Hook, E. B., M. S. Hsia, and R. R. Regal (2012). Accuracy of capture-recapture estimates of prevalence. *Epidemiologic Methods* 1(1), 1–11.
- Hook, E. B. and R. R. Regal (1999). Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of Clinical Epidemiology* 52(10), 917–926.
- Hook, E. B. and R. R. Regal (2000). Accuracy of alternative approaches to capture-recapture estimates of disease frequency: Internal validity analysis of data from five sources. *American Journal of Epidemiology* 152(8), 771–779.
- Hook, E. B., R. R. Regal, and R. Cormack (2000). On the need for a 16th and 17th recommendation for capture-recapture analysis. *Journal of Clinical Epidemiology* 53(12), 1275–1276.
- Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260), 663–685.

- Huberty, M., A. Serwaah, and G. Zachmann (2014). A flexible, scaleable approach to the international patent 'name game'. *Bruegel Working Paper* (September).
- Hutchinson, B., N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran (2022). Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1859–1876.
- Hutson, M. (2020). Core progress in AI has stalled in some fields. *Science* 368(6494), 927.
- Hwang, W. H. and R. Huggins (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika* 92(1), 229–233.
- Ibrahim, A., S. Rahnamayan, M. V. Martin, and K. Deb (2016). 3d-radvis: Visualization of pareto front in many-objective optimization. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 736–745. IEEE.
- ICH (2019). Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Technical report, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).
- ICH Steering Committee (2014). Final concept paper E9 (R1): Addendum to statistical principles for clinical trials on choosing appropriate estimands and defining sensitivity analyses in clinical trials dated 22 October 2014 endorsed by the ICH Steering Committee on 23 October 2014.
- Ilyas, I. F. and X. Chu (2019). *Data Cleaning*. New York, NY, USA: Association for Computing Machinery.
- International Labour Organization (2017a). *Global Estimates of Modern Slavery: Forced Labour and Forced Marriage*. Geneva.
- International Labour Organization (2017b). *Methodology of the Global Estimates of Modern Slavery: Forced Labour and Forced Marriage*. Geneva.
- Jain, V., T. Enamorado, and C. Rudin (2022). The importance of being earnest, ekundayo, or eswari: An interpretable machine learning approach to name-based ethnicity classification. *Harvard Data Science Review* 4(3). <https://hdsr.mitpress.mit.edu/pub/wgss79vu>.
- Jansche, M. (2007). A maximum expected utility framework for binary sequence labeling. In A. Zaenen and A. van den Bosch (Eds.), *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 736–743. Association for Computational Linguistics.

- Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* 84(406), 414–420.
- Johndrow, J. E., A. Bhattacharya, and D. B. Dunson (2017). Tensor decompositions and sparse log-linear models. *The Annals of Statistics* 45(1), 1 – 38.
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* 32(3), 241–254.
- Jutte, D. P., L. L. Roos, and M. D. Brownell (2011). Administrative record linkage as a tool for public health research. *Annual Review of Public Health* 32, 91–108.
- Kang, S., K. Gile, and M. Price (2020). Nested dirichlet process for population size estimation from multi-list recapture data. *arXiv preprint arxiv:2007.06160* (2016).
- Kim, K., M. Khabsa, and C. L. Giles (2016a). Inventor name disambiguation for a patent database using a random forest and DBSCAN. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2016-September*, 269–270.
- Kim, K., M. Khabsa, and C. L. Giles (2016b). Random forest DBSCAN for USPTO inventor name disambiguation. *arXiv:1602.01792*.
- Laake, J. L., D. S. Johnson, and P. B. Conn (2013). marked: An R package for maximum likelihood and markov chain monte carlo analysis of capture-recapture data. *Methods in Ecology and Evolution* 4(9), 885–890.
- Lalor, J. P., P. Rodriguez, J. Sedoc, and J. Hernandez-Orallo (2024). Item response theory for natural language processing. In M. Mesgar and S. Loáiciga (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, St. Julian’s, Malta, pp. 9–13. Association for Computational Linguistics.
- Landman, T. (2020). Measuring modern slavery: Law, human rights and new forms of data. *Human Rights Quarterly* 42(2), 303–331.
- Laohaprapanon, S., G. Sood, and B. Naji (2022). ethnicolr: Predict race and ethnicity from name. <https://github.com/appeler/ethnicolr>.
- Laplace, P.-S. (1820). *Théorie Analytique des Probabilités* (3 ed.). Paris, France.
- Li, G. C., R. Lai, A. D’Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu, and F. Lee (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010). *Research Policy* 43(6), 941–955.
- Liao, T., R. Taori, I. D. Raji, and L. Schmidt (2021). Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

- Lin, S., J. Hilton, and O. Evans (2021). TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Lincoln, F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. Technical report, United States Department of Agriculture.
- Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* 59(4), 1123–1130.
- Liu, Y., D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu (2023). GPTEval: NLG evaluation using GPT-4 with better human alignment. *arXiv Preprint arXiv:2303.16634*.
- Liu, Y., Z. Li, H. Xiong, X. Gao, and J. Wu (2010). Understanding of internal clustering validation measures. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 911–916.
- Lohr, S. L. (2021). *Sampling: design and analysis*. Chapman and Hall/CRC.
- Lum, K., J. Johndrow, and P. Ball (2015). dga: Capture-recapture estimation using Bayesian model averaging. <https://CRAN.R-project.org/package=dga>.
- Lum, K., M. E. Price, and D. Banks (2013). Applications of multiple systems estimation in human rights research. *American Statistician* 67(4), 191–200.
- Lyneham, S., C. Dowling, and S. Bricknell (2019). Estimating the dark figure of human trafficking and slavery victimisation in Australia. In *Crime and Justice Research 2019*. Australian Institute of Criminology.
- Madigan, D., J. York, and D. Allard (1995). Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique* 63(2), 215–232.
- Madigan, D. and J. C. York (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* 84(1), 19–31.
- Maidasani, H., G. Namata, B. Huang, and L. Getoor (2012). Entity resolution evaluation measures. Technical report, University of Maryland.
- Manrique-Vallier, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* 72(4), 1246–1254.
- Manrique-Vallier, D., P. Ball, and M. Sadinle (2021). Capture-recapture for casualty estimation and beyond: recent advances and research directions. *Statistics in the Public Interest: In Memory of Stephen E. Fienberg*, 15–31.
- Manrique-Vallier, D. and S. E. Fienberg (2008). Population size estimation using individual level mixture models. *Biometrical Journal* 50(6), 1051–1063.

- Manrique-Vallier, D., M. E. Price, and A. Gohdes (2013). Multiple systems estimation techniques for estimating casualties in armed conflicts. *Counting Civilian Casualties*, 165–181.
- Marchant, N. G. and B. I. Rubinstein (2017). In search of an entity resolution oasis: Optimal asymptotic sequential importance sampling. *Proceedings of the VLDB Endowment* 10(11), 1322–1333.
- Martínez-Plumed, F., S. Avin, M. Brundage, A. Dafoe, S. Ó. hÉigeartaigh, and J. Hernández-Orallo (2018). Between progress and potential impact of AI: the neglected dimensions. *arXiv preprint arXiv:1806.00610*.
- Martínez-Plumed, F., R. B. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence* 271, 18–42.
- Mathison, S. (2005). Evaluation. In S. Mathison (Ed.), *Encyclopedia of Evaluation*, pp. 140. Thousand Oaks, CA: Sage Publications, Inc.
- McVeigh, B. S., B. T. Spahn, and J. S. Murray (2019). Scaling Bayesian probabilistic record linkage with post-hoc blocking: An application to the california great registers. *arXiv:1905.05337*.
- Mehrbakhsh, B., F. Martínez-Plumed, and J. Hernández-Orallo (2023). Adversarial benchmark evaluation rectified by controlling for difficulty. In *ECAI 2023*, pp. 1696–1703. IOS Press.
- Mende, J. (2019). The concept of modern slavery: Definition, critique, and the human rights frame. *Human Rights Review* 20(2), 229–248.
- Menestrina, D., S. E. Whang, and H. Garciamolina (2010). Evaluating entity resolution results. *Proceedings of the VLDB Endowment* 3(1), 208–219.
- Michelson, M. and S. A. Macskassy (2009). Record linkage measures in an entity centric world. *Proceedings of the 4th Workshop on Evaluation Methods for Machine Learning*.
- Millard, G. and T. Blanchard (2022). Social data linkage environment (SDLE) methodology report: Linkage between the 2021 census of population file and the SDLE derived record depository. Statistics Canada Internal Report.
- Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229.

- Mizrahi, M., G. Kaplan, D. Malkin, R. Dror, D. Shahaf, and G. Stanovsky (2023). State of what art? a call for multi-prompt LLM evaluation. *arXiv preprint arXiv:2401.00595*.
- Monath, N., C. Jones, and S. Madhavan (2021). Patentsview: Disambiguating inventors, assignees, and locations. Technical report, American Institutes for Research, Arlington, Virginia.
- Monath, N., A. Kobren, A. Krishnamurthy, M. R. Glass, and A. McCallum (2019). Scalable hierarchical clustering with tree grafting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1438–1448.
- Montgomery, D. C. (2020). *Introduction to Statistical Quality Control*. John Wiley & Sons.
- Morrison, G., M. Riccaboni, and F. Pammolli (2017). Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Scientific Data* 4(1), 1–21.
- Müller, M.-C. (2017). Semantic author name disambiguation with word embeddings. In *Proceedings of the 21st International Conference on Theory and Practice of Digital Libraries*, pp. 300–311. Springer.
- Murphy, C., G. Kaiser, L. Hu, and L. Wu (2008). Properties of machine learning applications for use in metamorphic testing. *20th International Conference on Software Engineering and Knowledge Engineering, SEKE 2008*, 867–872.
- NannyML (2024). Estimation of performance of the monitored model. Accessed: May 22, 2024, https://github.com/NannyML/nannyml/blob/da33807cf7498ab9f8b8b924d2dba5814c2cf180/docs/how_it_works/performance_estimation.rst.
- Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959). Automatic linkage of vital records. *Science* 130(3381), 954–959.
- Oakden-Rayner, L., J. Dunnmon, G. Carneiro, and C. Re (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, pp. 151–159.
- Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*.
- Overstall, A. and R. King (2014). conting: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software* 58(7), 1–27.

- Papadakis, G., V. Efthymiou, E. Thanos, and O. Hassanzadeh (2021). Bipartite graph matching algorithms for clean-clean entity resolution: an empirical evaluation. *arXiv preprint arXiv:2112.14030*.
- Papadakis, G., E. Ioannou, E. Thanos, and T. Palpanas (2021). *The Four Generations of Entity Resolution*. Morgan & Claypool Publishers.
- Petersen, C. G. J. (1895). The yearly immigration of young Plaice into the Limfjord from the German Sea, etc. *Report of the Danish Biological Station to the Home Department 6*.
- Pezzoni, M., F. Lissoni, and G. Tarasconi (2014). How to kill inventors: Testing the massacrator© algorithm for inventor disambiguation. *Scientometrics 101*(1), 477–504.
- Phillips, A. and T. Clark (2021). Estimands in practice: Bridging the gap between study objectives and statistical analysis. *Pharmaceutical Statistics 20*(1), 68–76.
- Piper, N., M. Segrave, and R. Napier-Moore (2015). Editorial: What’s in a name? distinguishing forced labour, trafficking and slavery. *Anti-Trafficking Review* (5).
- Post, M. (2018). A call for clarity in reporting BLEU scores. *arXiv Preprint arXiv:1804.08771*.
- Poth, A., B. Meyer, P. Schlicht, and A. Riel (2020). Quality assurance for machine learning - an approach to function and system safeguarding. *Proceedings - 2020 IEEE 20th International Conference on Software Quality, Reliability, and Security, QRS 2020*, 22–29.
- Pushkarna, M., A. Zaldivar, and O. Kjartansson (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826.
- Qian, W., S. Abdelnasser, J. Oyarzun, A. Stelmack, I. Hekimi, and M. Mayda (2021). Quality measures for record linkage. Statistics Canada Internal Report.
- Quetelet, A. (1827). *Recherches sur la population, les naissances, les décès, les prisons, les dépôts de mendicité, etc. dans le royaume des Pays-Bas*. Bruxelles: chez H. Tarlier.
- Raji, I. D. and J. Buolamwini (2024). AI leaderboards are no longer useful. Accessed: 2024-05-31, <https://www.aisnakeoil.com/p/ai-leaderboards-are-no-longer-useful>.
- Ramanathan, A., L. L. Pullum, F. Hussain, D. Chakrabarty, and S. K. Jha (2016). Integrating symbolic and statistical methods for testing intelligent systems: Applications to machine learning and computer vision. *Proceedings of the 2016 Design, Automation and Test in Europe Conference and Exhibition, DATE 2016*, 786–791.

- Ramshaw, L. and R. E. Tarjan (2012). On minimum-cost assignments in unbalanced bipartite graphs. Technical report, HP Labs, Report HPL-2012-40R1.
- Regal, R. R. and E. B. Hook (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* 10(5), 717–721.
- Reich, Y. and S. Barai (1999). Evaluating machine learning models for engineering problems. *Artificial Intelligence in Engineering* 13(3), 257–272.
- Ridout, M. S. (2020). Discussion of read paper “Multiple-systems analysis for the quantification of modern slavery: Classical and Bayesian approaches”. *183*(3), 722.
- Rosenberg, A. and J. Hirschberg (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420.
- Saaty, T. L. (2003). Decision-making with the AHP: Why is the principal eigenvector necessary. *European Journal of Operational Research* 145(1), 85–91.
- Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association* 112(518), 600–612.
- Sadinle, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *The Annals of Applied Statistics* 12(2), 1013–1038.
- Sadinle, M. (2020). BRL: Beta record linkage. <https://CRAN.R-project.org/package=BRL>.
- Sanmartin, C., Y. Decady, R. Trudeau, A. Dasyuva, M. Tjepkema, P. Finés, R. Burnett, N. Ross, and D. G. Manuel (2016). Linking the Canadian community health survey and the Canadian mortality database: An enhanced data source for the study of mortality. In *Health Reports*, Volume 27 of *Catalogue no. 82-003-X*, pp. 1–11. Statistics Canada.
- Sariyar, M. and A. Borg (2022). RecordLinkage: Record linkage functions for linking and deduplicating data sets. <https://CRAN.R-project.org/package=RecordLinkage>.
- Särndal, C.-E., B. Swensson, and J. Wretman (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.
- Scarpa, S. (2020). *The Nebulous Definition of Slavery: Legal Versus Sociological Definitions of Slavery*, pp. 131–144. Cham: Springer International Publishing.

- Schelter, S., F. Biessmann, T. Januschowski, D. Salinas, S. Seufert, and G. Szarvas (2015). On challenges in machine learning model management. *IEEE Data Engineering Bulletin*.
- Schroff, F., D. Kalenichenko, and J. Philbin (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823.
- Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*. London: Charles Griffin.
- Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics* 42(2), 267–292.
- Seber, G. A. F. (1992). A review of estimating animal abundance ii. *International Statistical Review* 60(2), 129–166.
- Sekar, C. C. and W. E. Deming (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* 44(245), 101–115.
- Shi, Y., C. Otto, and A. K. Jain (2018). Face clustering: Representation and pairwise constraints. *IEEE Transactions on Information Forensics and Security* 13(7), 1626–1640.
- Sigmon, J. N. (2008). Combating modern-day slavery: Issues in identifying and assisting victims of human trafficking worldwide. *Victims and Offenders* 3(2-3), 245–257.
- Silverman, B. W. (2014). Modern slavery: An application of multiple systems estimation. *Home Office, London*. Available online: <https://www.gov.uk/government/publications/modern-slavery-an-application-of-multiple-systems-estimation>.
- Silverman, B. W. (2020). Multiple-systems analysis for the quantification of modern slavery: Classical and Bayesian approaches. *Journal of the Royal Statistical Society, Series A* 183, 691–736.
- Sjøberg, D. I. and G. R. Bergersen (2022). Construct validity in software engineering. *IEEE Transactions on Software Engineering* 49(3), 1374–1396.
- Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

- Statistics Canada (2017). *Record Linkage Project Process Model*. Catalog no. 12-605-X. Statistics Canada.
- Statistics Canada (2019). 2016 census of population coverage technical report. 98-303-X2016001.
- Steorts, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis* 10, 849–875.
- Steorts, R. C., R. Hall, and S. E. Fienberg (2016). A Bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association* 111, 1648–1659.
- Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press.
- Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2), 111–133.
- Streamlit (2023). Streamlit - the fastest way to build data apps in python. <https://github.com/streamlit/streamlit>.
- Tam, D., N. Monath, A. Kobren, A. Traylor, R. Das, and A. McCallum (2019). Optimal transport-based alignment of learned character representations for string similarity. *arXiv:1907.10165*.
- Tancredi, A. and B. Liseo (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics* 5, 1553 – 1585.
- Tang, J., J. P. Reiter, and R. C. Steorts (2020). Bayesian modeling for simultaneous regression and record linkage. In J. Domingo-Ferrer and K. Muralidhar (Eds.), *Privacy in Statistical Databases*, pp. 209 – 223. Lecture Notes in Computer Science 12276, Cham, Switzerland: Springer.
- Tong, R. M., N. D. Newman, G. Berg-Cross, and F. Rook (1987). Performance evaluation of artificial intelligence systems. Technical report.
- Toole, A., C. Jones, and S. Madhavan (2021). Patentsview: An open data platform to advance science and technology policy.
- Trajtenberg, M. and G. Shiff (2008). *Identification and Mobility of Israeli Patenting Inventors*. Pinhas Sapir.
- Traylor, A., N. Monath, R. Das, and A. McCallum (2017). Learning string alignments for entity aliases. In *Proceedings of the 31st Conference on Neural Information Processing Systems*.

- Triantaphyllou, E., B. Shu, S. N. Sanchez, and T. Ray (1998). Multi-criteria decision making: An operations research approach. *Encyclopedia of Electrical and Electronics Engineering* 15(1998), 175–186.
- Tuncali, C. E., G. Fainekos, D. Prokhorov, H. Ito, and J. Kapinski (2020). Requirements-driven test generation for autonomous vehicles with machine learning components. *IEEE Transactions on Intelligent Vehicles* 5(2), 265–280.
- UK Parliament (2015). Modern slavery act 2015. Available online: http://www.legislation.gov.uk/ukpga/2015/30/pdfs/ukpga_20150030_en.pdf.
- United Kingdom Human Trafficking Centre (2014). A strategic assessment on the nature and scale of human trafficking in 2013. Technical report, National Crime Agency.
- United Nations (2000). Protocol to prevent, suppress and punish trafficking in persons, especially women and children, supplementing the United Nations convention against transnational organized crime.
- United Nations (2001). United Nations convention against transnational organized crime (A/RES/55/25).
- UNODC (2018a). Monitoring target 16.2 of the United Nations sustainable development goals: Multiple systems estimation of the numbers of presumed victims of trafficking in persons: Ireland.
- UNODC (2018b). Monitoring target 16.2 of the United Nations sustainable development goals: Multiple systems estimation of the numbers of presumed victims of trafficking in persons: Romania.
- UNODC (2018c). Monitoring target 16.2 of the United Nations sustainable development goals: Multiple systems estimation of the numbers of presumed victims of trafficking in persons: Serbia.
- USPTO (2023). Data download tables. Accessed: November 16, 2023, <https://patentsview.org/download/data-download-tables>.
- van der Laan, J. and B. Bakker (2015). Indicator for the representativeness of linked sources. In *Proceedings of the Conference on New Techniques and Technologies for Statistics*.
- van Dijk, J., M. Cruyff, P. G. M. van der Heijden, and S. L. J. Kragten-Heerdink (2017). Monitoring target 16.2 of the United Nations sustainable development goals: A multiple systems estimation of the numbers of presumed human trafficking victims in the netherlands in 2010-2015 by year, age, gender, form of exploitation and nationality. Technical report, United Nations Office on Drugs and Crime.

- Ventura, S. L., R. Nugent, and E. R. Fuchs (2013). Methods matter: Rethinking inventor disambiguation with classification & labeled inventor records. In *Academy of Management Proceedings*, Volume 2013. Academy of Management Briarcliff Manor, NY 10510.
- Ventura, S. L., R. Nugent, and E. R. Fuchs (2015). Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy* 44 (9), 1672–1701.
- Vincent, K., K. Bales, D. P. Durgana, M. Cruyff, P. G. van der Heijden, and J. van Dijk (2020a). Misunderstandings of multiple systems estimation: A response to “on the unreliability of multiple systems estimation for estimating the number of potential victims of modern slavery in the UK” by whitehead, jackson, balch, and francis (2019). *Journal of Human Trafficking*, 1–6.
- Vincent, K., K. Bales, D. P. Durgana, M. Cruyff, P. G. van der Heijden, and J. van Dijk (2020b). Vincent et al. concluding response to whitehead et al. *Journal of Human Trafficking*, 1–2.
- Walk Free Foundation (2013). *The Global Slavery Index*. <https://www.walkfree.org/global-slavery-index/>.
- Wang, T., H. Lin, C. Fu, X. Han, L. Sun, F. Xiong, H. Chen, M. Lu, and X. Zhu (2022). Bridging the gap between reality and ideality of entity matching: A revisiting and benchmark re-construction. arxiv:2205.05889.
- Wang, X., L. Jiang, J. Hernandez-Orallo, L. Sun, D. Stillwell, F. Luo, and X. Xie (2023). Evaluating general-purpose AI with psychometrics. *arXiv preprint arXiv:2310.16379*.
- Wanzer, D. L. (2021). What is evaluation?: Perspectives of how evaluation differs (or not) from research. *American Journal of Evaluation* 42(1), 28–46.
- Whitehead, J., J. Jackson, A. Balch, and B. Francis (2019). On the unreliability of multiple systems estimation for estimating the number of potential victims of modern slavery in the UK. *Journal of Human Trafficking*, 1–13.
- Whitehead, J., J. Jackson, A. Balch, and B. Francis (2020). Whitehead et al. response to “misunderstandings of multiple systems estimation.”. *Journal of Human Trafficking*, 1–5.
- Winkler, W. E. and Y. Thibaudeau (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census. Technical report, United States Bureau of the Census, Working Paper Number RR91-09.

- Wittes, J. T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association* 69(345), 93–97.
- Worthington, H., R. McCrea, R. King, and K. S. Vincent (2021). How ideas from ecological capture-recapture models may inform multiple systems estimation analyses. *Crime & Delinquency* 67(13-14), 2278–2294.
- Wortman, J. P. H. (2019). *Record Linkage Methods with Applications to Causal Inference and Election Voting Data*. Ph. D. thesis, Department of Statistical Science, Duke University.
- Xia, C., C. Xing, J. Du, X. Yang, Y. Feng, R. Xu, W. Yin, and C. Xiong (2024). FOFO: A benchmark to evaluate LLMs’ format-following capability. *arXiv preprint arXiv:2402.18667*.
- Xie, J., S. Kelley, and B. K. Szymanski (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)* 45(4), 1–35.
- Yang, G.-C., C. Liang, Z. Jing, D.-R. Wang, and H.-C. Zhang (2017). A mixture record linkage approach for US patent inventor disambiguation. In *Advanced Multimedia and Ubiquitous Engineering*, Volume 448, pp. 331–338. Springer.
- Yao, Y., A. Vehtari, and A. Gelman (2020). Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. *arXiv preprint arXiv:2006.12335*.
- Yin, D., K. Motohashi, and J. Dang (2020). Large-scale name disambiguation of chinese patent inventors (1985–2016). *Scientometrics* 122(2), 765–790.
- Yip, S. F., M. Richard, S. E. Fienberg, B. W. Junker, R. E. Laporte, and I. M. Libman (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology* 142(10), 1047–1058.
- Yip, S. F., N. Tajima, G. A. F. Seber, S. T. Buckland, R. M. Cormack, N. Unwin, Y.-F. Chang, S. E. Fienberg, B. W. Junker, R. E. LaPorte, I. M. Libman, and D. J. McCarty (1995). Capture-recapture and multiple-record systems estimation II: Applications in human diseases. *American Journal of Epidemiology* 142(10), 1059–1068.
- York, J. C. and D. Madigan (1992). Bayesian methods for estimation of the size of a closed population. Technical Report 234, University of Washington, Seattle, Wa.
- Zhang, J. M., M. Harman, L. Ma, and Y. Liu (2022). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 48(1), 1–37.

- Zhang, X., B. Yu, H. Yu, Y. Lv, T. Liu, F. Huang, H. Xu, and Y. Li (2023). Wider and deeper LLM networks are fairer LLM evaluators. *arXiv preprint arXiv:2308.01862*.
- Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. (2024). Judging LLM-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36.
- Zhou, K., Y. Zhu, Z. Chen, W. Chen, W. X. Zhao, X. Chen, Y. Lin, J.-R. Wen, and J. Han (2023). Don't make your LLM an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.