Record Linkage Methods with Applications to Causal Inference and Election Voting Data

by

Joan Pearson Heck Wortman

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Fan Li

_____
Rebecca Steorts

_____
Sunshine Hillygus

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
2019

Abstract

Record Linkage Methods with Applications to Causal Inference and Election Voting Data

by

Joan Pearson Heck Wortman

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Fan Li

_____
Rebecca Steorts

_____
Sunshine Hillygus

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
2019

Abstract

Probabilistic record linkage enables researchers and analysts to combine data from multiple data sources to conduct statistical analysis. This analysis may be to answer causal questions, to predict future outcomes, or to provide descriptive statistics. In this dissertation, I develop methodology for probabilistic record linkage for two scenarios: general causal inference applications with linked data, and identifying previously removed voters in North Carolina who cast provisional ballots in 2016.

In Chapter 2, we develop methodology for causal inference in observational studies when using propensity score subclassification on data constructed with probabilistic record linkage techniques. We focus on scenarios where covariates and binary treatment assignments are in one file and outcomes are in another file, and the goal is to estimate an additive treatment effect by merging the files. We assume that the files can be linked using variables common to both files, e.g., names or birth dates, but that links are subject to errors, e.g., due to reporting errors in the linking variables. We develop methodology for cases where such reporting errors are independent of the other variables on the files. We describe conceptually how linkage errors can affect causal estimates in subclassification contexts. We also present and evaluate several algorithms for deciding which record pairs to use in estimation of causal effects. Using simulation studies, we demonstrate that case selection procedures can result in improved accuracy in estimates of treatment effects from linked data compared to using only cases known to be true links.

In Chapter 3, we introduce a model for Bayesian record linkage and clustered submodels, which we call BRACS. The model is designed for combining two sets of data in which there are differences in the comparison distributions for links and non-links, conditional on attributes observed in one of the files. We use simulation studies to demonstrate that the proposed approach can yield improvements in classifying record pairs as links versus non-links.

In Chapter 4, we apply BRACS to 2016 voting data from North Carolina. We describe the process of provisional voting and the list of provisional voters provided by the North Carolina Board of Elections. We provide background on the North Carolina voter file, of which we use a snapshot from November 2016. We outline the limitations of exact-matching the two files using only the state-provided identifiers. Finally, we use BRACS to link the two files, with and without the state-provided identifiers, in order to estimate the number of removed voters who cast provisional ballots in the November 2016 election in North Carolina.

In Chapter 5, we modify BRACS to relax the assumption of conditionally independent field comparisons, motivated by the correlation between party registration and race in North Carolina. We outline a method for accounting for this correlation, in which we combine two dependent comparison fields into one joint comparison field. We use simulation studies to demonstrate that this can yield improvements in linkage quality, and we also outline when it may not be appropriate to use. Finally, we apply the results to the data in Chapter 4 and re-estimate the number of removed voters with the joint-comparison BRACS.

To Zack, Marmy, and Father. And to Sally, the perfect motivation to cross the finish line.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I'd like to thank my advisor Jerry Reiter, who has not only taught me how to be a researcher and statistician but has also been an invaluable mentor throughout the past four years. Thank you also to Beka Steorts, Fan Li, and Sunshine Hillygus on my dissertation committee for their feedback and guidance. Thank you to Mauricio Sadinle for his feedback on the BRACS model.

This research was supported by NSF Grant SES 1131897 and I am grateful for the opportunity.

Thank you to Lori Rauch who has, with constant kindness and patience, guided and supported me through all the administrative hurdles of being a student, taking leave, and being employed while finishing my degree. I've learned so much in the Duke Statistical Science Department and would like to thank all the faculty and fellow students who have made the past few years enriching, fun, and exciting. My colleagues and friends at the Democratic National Committee have been kind, supportive, and incredibly accommodating throughout the past year I've been employed there. Special thanks to Catherine Tarsney for her friendship and support.

Finally, I'd like to thank my family–my parents and siblings, and above all my husband Zack, who has supported me in my statistics journey from my first class freshman year of college to the final stretch of the dissertation.

1

Introduction

Record linkage enables researchers and analysts to combine data from multiple data sources to conduct analysis. This analysis may be to answer causal questions, to predict future outcomes, or to provide descriptive statistics. In the absence of unique identifiers, researchers often rely on indirect identifiers such as names, addresses, and demographic information. Probabilistic record linkage (Fellegi and Sunter, 1969; Tancredi and Liseo, 2011; Dalzell and Reiter, 2018; Herzog et al., 2007) allows for the use of this imprecise information in a way that acknowledges the uncertainty associated with imperfect linkage. In this dissertation, I develop methodology for probabilistic record linkage for two scenarios: general causal inference applications with linked data, and identifying previously removed voters in North Carolina who cast provisional ballots in 2016. The thesis is organized as follows.

In Chapter 2, we develop methodology for causal inference in observational studies when using propensity score subclassification on data constructed with probabilistic record linkage techniques. We focus on scenarios where covariates and binary treatment assignments are in one file and outcomes are in another file, and the goal is to estimate an additive treatment effect by merging the files. We assume that the files

can be linked using variables common to both files, e.g., names or birth dates, but that links are subject to errors, e.g., due to reporting errors in the linking variables. We develop methodology for cases where such reporting errors are independent of the other variables on the files. We describe conceptually how linkage errors can affect causal estimates in subclassification contexts. We also present and evaluate several algorithms for deciding which record pairs to use in estimation of causal effects. Using simulation studies, we demonstrate that case selection procedures can result in improved accuracy in estimates of treatment effects from linked data compared to using only cases known to be true links.

In Chapter 3, we introduce a model for Bayesian record linkage and clustered sub-models, which we call BRACS. The model is designed for combining two sets of data in which there are differences in the comparison distributions for links and non-links, conditional on attributes observed in one of the files. We introduce notation and motivate modification of existing Bayesian approaches to record linkage, and then outline the model. We use simulation studies to demonstrate that the proposed approach can yield improvements in classifying record pairs as links versus non-links.

In Chapter 4, we apply BRACS to 2016 voting data from North Carolina. We describe the process of provisional voting and the list of provisional voters provided by the North Carolina Board of Elections. We provide background on the North Carolina voter file, of which we use a snapshot from November 2016. We outline the limitations of exact-matching the two files using only the state-provided identifiers. Finally, we use BRACS to link the two files, with and without the state-provided identifiers, in order to estimate the number of removed voters who cast provisional ballots in the November 2016 election in North Carolina.

In Chapter 5, we modify BRACS to relax the assumption of conditionally independent field comparisons, motivated by the correlation between party registration and race in North Carolina. We outline a method for accounting for this correlation,

in which we combine two dependent comparison fields into one joint comparison field. We use simulation studies to demonstrate that this can yield improvements in linkage quality, but we also outline when it may not be appropriate to use. Finally, we apply the results to the data in Chapter 4 and re-estimate the number of removed voters with the joint-comparison BRACS.

Each chapter includes the background information, citations, and notation relevant to the topic covered. I note that the material in Chapter 2 is similar to Wortman and Reiter (2018).

2

Simultaneous Record Linkage and Causal Inference Using Propensity Score Stratification

2.1 Introduction

Increasingly, researchers are linking data collected in planned studies to data available in administrative sources, such as electronic health records and Medicare claims data, in order to enhance analyses of causal questions. For example, linking can enable researchers to evaluate long-term outcomes, as well as outcomes not measured in the planned study, without expensive de novo primary data collection. It also can allow researchers to incorporate important covariates not collected in the planned study, thereby reducing effects of unmeasured confounding and facilitating more nuanced estimation of treatment effects.

When perfectly measured unique identifiers, such as Medicare patient IDs or social security numbers, are available on both files, the linkage is a relatively straightforward task: one simply merges on the identifiers and proceeds with statistical inference. In many settings, however, such identifiers are unavailable on at least one file, e.g., because of privacy restrictions, and record linkage must be based on indirect

identifiers like birth dates, diagnosis codes, demographic characteristics, and names that could differ on the files for the same individual. In such contexts, typical record linkage procedures involve scoring potentially linked record pairs based on similarity of the linking fields—where larger values imply more confidence in the correctness of the proposed link—and selecting as links those pairs whose score exceeds some threshold (Fellegi and Sunter, 1969; Winkler, 1990; Jin et al., 2003; Herzog et al., 2007; Christen, 2012).

Methodologies for record linkage with indirect identifiers and for causal inference in observational data are well established; however, we are not aware of methodology developed specifically for causal inference with linked observational data. Yet, the fact that we seek causal inferences clearly affects the consequences of incorrect linkages. For example, suppose that a researcher has some File A that contains treatment and covariate values for a set of patients, and some File B that contains long term outcomes for these and other patients. The researcher uses propensity score matching (Rosenbaum and Rubin, 1983, 1984) to create balanced treatment and control groups from File A. In this case, incorrect linkages for records excluded from the matched control set do not affect the causal estimates, whereas incorrect links for those in the treated and matched control sets do. This example suggests general questions. When estimating a causal effect, should we use only linked pairs where the link has near certain probability of being correct, or can we benefit from allowing lower probability links to enter the causal estimate? If the latter, how do we draw the line on what to include and exclude? As far as we can tell, these questions have not been addressed in the literature.

In this chapter, we begin to address these questions. Specifically, we present and evaluate several algorithms for estimation of additive treatment effects when using subclassification on propensity scores with inexactly linked data. We consider observational studies where File A includes a binary treatment and covariate val-

ues, and File B includes outcome values. We develop algorithms assuming that the processes generating mismatches in the linking variables across files are unrelated to other variables on the files; for example, errors in the names or birth dates in the two files are independent of the outcomes, treatments, and causally relevant covariates. Throughout, we assume both the stable unit treatment value assumption (SUTVA) (Rubin, 1978, 1980, 1990) and strong ignorability (Rubin, 1978; Rosenbaum and Rubin, 1984). SUTVA requires that one unit's treatment status does not affect another unit's potential outcomes and also that there are no hidden levels of treatment. Strong ignorability requires that all units have a non-zero probability of being in the treatment and control groups, and that treatment assignment depends only on observed covariates.

The basic strategy underpinning the different algorithms is as follows. First, we order the pairs selected by the record linkage procedure from highest to lowest linking scores. Second, we peel off the cases deemed to represent correct links with near certainty. Third, starting from these certainty cases, we sequentially concatenate new linked records to augment the set of cases that could be used to estimate treatment effects, each time computing some criterion intended to increase when adding inexact matches. Finally, we find the set of cases that corresponds to the smallest value of the criterion, and use this set of records in the causal inference. As we demonstrate in simulation studies, case selection procedures following this strategy can reduce mean squared errors compared to using only the certainty cases or using more liberal thresholds.

The remainder of the chapter is organized as follows. In Section 2.2, we provide background on subclassification on propensity scores and on threshold based record linkage techniques. In Section 2.3, we discuss the effects of linkage errors on causal inferences when using propensity score subclassification. In Section 2.4, we describe several algorithms for choosing record pairs. In Section 2.5, we present simulation re-

sults that compare the different algorithms and illustrate their potential benefits and limitations. In Section 2.6, we summarize the findings and suggest future research topics.

2.2 Background

Setting aside complexities associated with three or more possible treatments, which we leave for future consideration, we focus on scenarios where there is a binary treatment. Let $w_i = 1$ indicate that individual i is assigned treatment, and $w_i = 0$ indicate that individual i is assigned control. Let x_i indicate a $p \times 1$ vector of causally relevant covariates for individual i. Let $Y_i(1)$ be the value of the outcome for individual i when $w_i = 1$, and $Y_i(0)$ be the value of the outcome for individual i when $w_i = 0$. Let $\tau_i = Y_i(1) - Y_i(0)$ be the treatment effect for individual i. Throughout, we assume additive treatment effects, that is, $\tau_i = \tau$ for all individuals i. Finally, let $y_i = w_i Y_i(1) + (1 - w_i) Y_i(0)$ be the observed outcome for any individual i.

We reserve the term link (and its derivatives) for when some record in File A and some record in File B are deemed to belong to the same individual, and the term match for operations involving balancing covariate distributions in treated and control groups.

2.2.1 Propensity scores and subclassification

The propensity score is defined as $e(x) = P(w = 1|x)$, i.e., the probability of being assigned treatment given covariate pattern x. It can be shown that treatment assignment is independent of x given $e(x)$. Thus, treated and control units with the same propensity score have the same distribution of x, so that analysts who compare treated and control units with the same propensity score effectively ensure that x does not confound estimation of the treatment effects (Rosenbaum and Rubin, 1984).

Given sets of individuals assigned to treatment and control, analysts can estimate each individual's $e(x_i)$ using binary regression techniques, such as logistic regression, where the outcome is treatment status and the predictors are the relevant covariates. Propensity scores are used in a variety of ways in causal inference, (Stuart, 2010; Imbens and Rubin, 2015) including matching, inverse probability weighting, and subclassification as we do here.

In propensity score subclassification, the goal is to partition the collected data into J strata, called subclasses, in which treated and control units have similar covariate distributions. The partition often is based on equally spaced quantiles of the propensity scores, e.g., every twentieth percentile. Analysts manually adjust the breaks as necessary to ensure sufficient sample sizes or improve covariate balance in each subclass. In the simulations, we use the common choice of $J = 5$ and breaks based on manual specifications of propensity score quantiles.

Let $j \in \{1, \ldots, J\}$ index the J subclasses, and let $\mathcal{S}_j$ where $j = 1, \ldots, J$ represent the set of individuals in subclass j. For each j, let n_{1j} and n_{0j} be the number of individuals in $\mathcal{S}_j$ with $w_i = 1$ and $w_i = 0$, respectively. Let $\bar{y}_{1j} = \sum_{i \in \mathcal{S}_j} w_i y_i / n_{1j}$ and $\bar{y}_{0j} = \sum_{i \in \mathcal{S}_j} (1 - w_i) y_i / n_{0j}$. Within each subclass $j = 1, \ldots, J$, we compute the estimated subclass average treatment effect, $\hat{\tau}_j = \bar{y}_{1j} - \bar{y}_{0j}$. We estimate τ using the weighted average,

$$\hat{\tau} = \sum_{j=1}^{J} \lambda_j \hat{\tau}_j. \tag{2.1}$$

A typical value of λ_j, which we use in the simulations, is $\lambda_j = \frac{n_j}{n}$, where $n_j = n_{1j} + n_{0j}$ and $n = \sum_j n_j$.

For the estimated variances, it is common to use

$$\hat{var}(\hat{\tau}) = \sum_{j=1}^{J} \lambda_j^2 \hat{var}(\hat{\tau}_j) = \sum_{j=1}^{J} \lambda_j^2 \left(\frac{s_{0j}^2}{n_{0j}} + \frac{s_{1j}^2}{n_{1j}} \right), \tag{2.2}$$

where $s_{0j}^2 = \sum_{i \in \mathcal{S}_j} (y_i(1-w_i) - \bar{y}_{0j})^2/(n_{0j}-1)$ and $s_{1j}^2 = \sum_{i \in \mathcal{S}_j} (y_i w_i - \bar{y}_{1j})^2/(n_{1j}-1)$. We note that this variance estimator is biased for the true variance of (2.1), as it does not account for estimation of the propensity scores (Williamson et al., 2012).

Residual imbalance often remains after subclassification. To reduce the effects of the remaining imbalance, analysts can regress y on w and some subset of x within the subclasses (Rosenbaum and Rubin, 1984; D'Agostino, 1998). Let $\hat{\beta}_j$ be the estimated coefficient of the indicator for w in the regression in subclass j. To estimate τ, we can use $\hat{\tau}_\beta = \sum_{j=1}^{J} \lambda_j \hat{\beta}_j$. We can estimate the variance of $\hat{\tau}_\beta$ using (2.2), replacing the two-sample variance estimator with the estimated variance of $\hat{\beta}_j$ from each within-subclass regression.

2.2.2 Record Linkage

We consider scenarios where an analyst seeks to link two files, File A comprising n_A records and File B comprising n_B records, using imperfect linking variables present in both files. In such settings, many analysts use the probabilistic record linkage framework formalized by Fellegi and Sunter (1969). For all possible record pairs (i, i'), where record i is in File A and record i' is in File B, the analyst computes some measure $S(\gamma_{ii'})$ that reflects the similarity of the linking variables for record i from File A to those for record i' from File B. Record pairs with similarity scores above an analyst-specified threshold are declared links, and others are declared either non-links or uncertain status. Uncertain links can be sent to clerical review for adjudication or, as is often done, treated as non-links as we do here.

More precisely, suppose that we have F linking variables; these often are called fields in the record linkage literature. For each field $f \in (1, \ldots, F)$, let $\gamma_{fii'}$ be a score reflecting the similarity in field f for that pair. Typically, we set $\gamma_{fii'} = 1$ when the values of field f for records i and i' are identical or within some acceptable tolerance, and set $\gamma_{fii'} = 0$ otherwise. For each record pair (i, i'), let $\gamma_{ii'} = (\gamma_{1ii'}, \ldots, \gamma_{Fii'})$ be

the vector comprising the comparisons for each linking field. Following Fellegi and Sunter, (1969) we assume that $\gamma_{ii'}$ is a random realization from a mixture of two distributions, one for true links and one for non-links. Let $\mathcal{M}$ be the set of true links in File A and File B, and let $\mathcal{U}$ be the set of non-links in these files. The mixture model for $\gamma_{ii'}$ is thus

$$\gamma_{ii'} \mid (i,i') \in \mathcal{M} \quad \sim \quad f(\theta_m) \tag{2.3}$$

$$\gamma_{ii'} \mid (i,i') \in \mathcal{U} \quad \sim \quad f(\theta_u), \tag{2.4}$$

where θ_m and θ_u are parameters specific to each class. For computational simplicity, usually one assumes conditional independence of the $\gamma_{fii'}$ both across fields and pairs, computing

$$m(\gamma_{ii'}) \quad = \quad P(\gamma_{ii'} \mid \theta_m, (i,i') \in \mathcal{M}) = \prod_f P(\gamma_{fii'} \mid \theta_{mf}, (i,i') \in \mathcal{M}) = \prod_f \theta_{mf}^{\gamma_{fii'}} (1-\theta_{mf})^{1-\gamma_{fii'}}, \tag{2.5}$$

$$u(\gamma_{ii'}) \quad = \quad P(\gamma_{ii'} \mid \theta_u, (i,i') \in \mathcal{U}) = \prod_f P(\gamma_{fii'} \mid \theta_{uf}, (i,i') \in \mathcal{U}) = \prod_f \theta_{uf}^{\gamma_{fii'}} (1-\theta_{uf})^{1-\gamma_{fii'}}. \tag{2.6}$$

Fellegi and Sunter (1969) use a decision-theoretic approach to minimize Type I and Type II error rates, that is, erroneously linking or erroneously not linking records, respectively. They compute the likelihood ratio, $R(i,i') = \frac{m(\gamma_{ii'})}{u(\gamma_{ii'})}$. Values of $R(i,i')$ above some upper threshold are deemed links, and values below some lower threshold are deemed non-links. When all linking fields are binary and one assumes conditional independence, it is common to write $R(i,i')$ as

$$S(\gamma_{ii'}) = \sum_{f=1}^{F} log_2 \left(\frac{\theta_{mf}}{\theta_{uf}} \right) \gamma_{fii'} + log_2 \left(\frac{1-\theta_{mf}}{1-\theta_{uf}} \right) (1-\gamma_{fii'}). \tag{2.7}$$

$S(\gamma_{ii'})$ is often called the linking score for pair (i,i').

String data, including names, complicate the construction and computation of similarity scores (Newcombe et al., 1959; Jaro, 1989, 1995; Larsen and Rubin, 2001).

A common approach, which we use here and now review briefly, is to compute Jaro-Winkler scores (Winkler, 1990) for the string fields. Suppose we seek to compare two strings on a set of characters, where the string in File A has d such characters and the string in File B has r such characters. Suppose the two strings have $c > 0$ of these characters in common and t characters that are transposed. Suppose that we assign a weight to each string, say W_A and W_B, as well as a weight to transpositions, say W_t. Then, the Jaro score is $\Phi_J = W_A(c/d)+W_B(c/r)+W_t(c-t)/c$. The Jaro-Winkler score boosts the weight of agreement early in a string, resulting in $\Phi_{JW} = \Phi_J+0.1g(1-\Phi_J)$, where g is the number of characters among the first four that agree in the two strings. Scores range from 0 (no agreement) to 1 (full agreement), and can be easily estimated using the "jarowinkler" function in the "RecordLinkage" package in R. Analysts can use the values of Φ_{JW} as similarity measures, or turn each $\Phi_{JW}(i, i')$ into a binary $\gamma_{fii'}$ by setting $\gamma_{fii'} = 1$ when $\Phi_{JW}(i, i') > t_o$ and $\gamma_{fii'} = 0$ otherwise. In the simulations of Section 2.5, we use this approach with $t_o = 0.95$.

The linkage process can be subject to errors, e.g., records belonging to two different individuals are linked, or incompleteness, e.g., an individual with a record in File A truly does not have a record in File B. It is well known that incorrect and incomplete linkages can degrade the quality of subsequent statistical inferences (Herzog et al., 2007; Gu et al., 2003; Jaro, 1989; Belin and Rubin, 1995). There has been some work on accounting for such errors in inferences for regression modeling (Scheuren and Winkler, 1997; Lahiri and Larsen, 2005; Chambers, 2008; Kim and Chambers, 2009, 2011; Chipperfield et al., 2011; Gutman et al., 2013; Dalzell and Reiter, 2018). We are not aware of propensity score methods for causal inference that explicitly account for inexact linkage.

2.3 The effects of linkage errors on causal estimands

In this section, we provide intuition on the impacts of linkage errors on causal estimates made with propensity score subclassification. The discussion is organized around Figure 2.1, which highlights four types of linkage errors in the context of subclassification. Each terminal node in the tree represents a possible outcome of the linkage process. In each node, the listed value of $E(\hat{\tau}^*)$ is the expected treatment effect for the extreme case that every link falls into that node. It offers a sense of the contributions to bias in treatment effect estimation that one can expect from different types of incorrect linkages. μ_1 and μ_0 are population marginal means of the outcomes for treated and control units, respectively. Throughout we assume that it is possible to balance covariate distributions in the treatment and control groups with subclassification on properly linked records.

As we link records with (x_i, w_i) measured in File A to records with $y_{i'}$ measured in File B, causal estimates are based on values that may differ from the true values due to linkage errors. For each record i in File A and its linked record i' in File B, let $y_i^* = y_i$ when the linked pair is correct, i.e., records i and i' belong to the same individual, and let $y_i^* = y_{i'}$ when the linked pair is incorrect. Quantities from Section 2.2.1 use y_i^* rather than y_i, so that, for example, the within-class estimate of treatment effect is

$$\hat{\tau}_j^* = \bar{y}_{1j}^* - \bar{y}_{0j}^* = \sum_{i \in \mathcal{S}_j} w_i y_i^* / n_{1j} - \sum_{i \in \mathcal{S}_j} (1 - w_i) y_i^* / n_{0j}. \tag{2.8}$$

Inferences then are based on

$$\hat{\tau}^* = \sum_{j=1}^{J} \lambda_j \hat{\tau}_j^* \tag{2.9}$$

$$\hat{var}(\hat{\tau}^*) = \sum_{j=1}^{J} \lambda_j^2 \left(\frac{s_{0j}^{2*}}{n_{0j}} + \frac{s_{1j}^{2*}}{n_{1j}} \right), \tag{2.10}$$

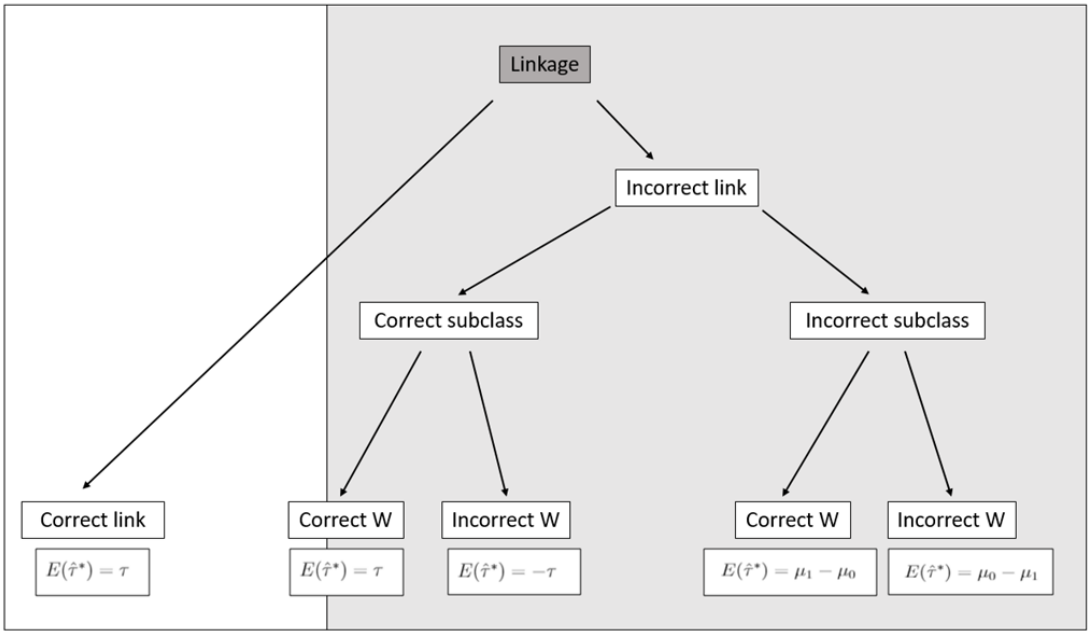

FIGURE 2.1: Graphical representation of the effects of linkage errors on estimation of treatment effects in propensity score subclassification.

where $s^{2*}_{0j} = \sum_{i \in \mathcal{S}_j} (y^*_i(1-w_i) - \bar{y}^*_{0j})^2/(n_{0j}-1)$ and $s^{2*}_{1j} = \sum_{i \in \mathcal{S}_j} (y^*_i w_i - \bar{y}^*_{1j})^2/(n_{1j}-1)$. As subclassification is based only on the (x_i, w_i) from records in File A, values of the propensity score estimation need not be affected by the record linkage. Typically, the linked file used in treatment effect estimation does not include all n_A records from File A, in which case $\mathcal{S}_j$ should be interpreted as restricted to the set of linked pairs used in analysis, with (n_{0j}, n_{1j}) computed over that restricted set.

2.3.1 Matching subclass and treatment status

It is possible to link two records that do not belong to the same individual yet not incur bias if (i) the two records' actual covariate values place them into a common subclass and (ii) they experienced the same treatment assignment. To illustrate, suppose the record linkage algorithm links record i in File A named David Copperfield to record i' in File B named Davy Copperfull. These records are not a true link, but they have similar background information. David is 43 years old, and Davy is

42. Both are men with 3 children. Both were assigned to treatment, say a new drug intended to reduce high blood pressure. If the propensity score model conditions on age, sex, and number of children, the subclassification algorithm may well put them in the same subclass j.

When using (2.9) to estimate treatment effects, linkage errors like this tend to have only modest impacts. Suppose, for example, that this type of linkage error occurs only for two treated individuals i and k in subclass j such that, after linkage, $y_i^* = y_k = Y_k(1)$ and $y_k^* = y_i = Y_i(1)$. In this case, $\hat{\tau}^* = \hat{\tau}$, so there is no effect on the differences in means inference. Of course, the regression-adjusted treatment effect estimate changes unless $x_i = x_k$.

More formally, suppose again for simplicity that only one record k in subclass j is incorrectly linked to a record k' with the same w and subclass. We consider treatment assignment within the subclass as completely randomized (Rosenbaum, 1995), that is, consider $e(x_i) = n_{1j}/n_j$ for all $i \in \mathcal{S}_j$. In this case, averaging over the treatment assignment, we have

$$\begin{aligned}
E(\hat{\tau}_j^*) &= E\left(\sum_{i\in\mathcal{S}_j} w_i y_i^*/n_{1j} - \sum_{i\in\mathcal{S}_j}(1-w_i)y_i^*/n_{0j}\right) \\
&= E\left(\sum_{i\in\mathcal{S}_j, i\neq k'} w_i y_i/n_{1j} - \sum_{i\in\mathcal{S}_j, i\neq k'}(1-w_i)y_i/n_{0j} + w_{k'}y_{k'}/n_{1j} - (1-w_{k'})y_{k'}/n_{0j})\right) \\
&= \sum_{i\in\mathcal{S}_j, i\neq k'}\left(\frac{n_{1j}/n_j}{n_{1j}}Y_i(1) - \frac{1-n_{1j}/n_j}{n_{0j}}Y_i(0)\right) + \frac{n_{1j}/n_j}{n_{1j}}Y_{k'}(1) - \frac{1-n_{1j}/n_j}{n_{0j}}Y_{k'}(0) \\
&= (1/n_j)\left(\sum_{i\in\mathcal{S}_j, i\neq k'}\tau_i + \tau_{k'}\right). \qquad (2.11)
\end{aligned}$$

With additive treatment effects, this expectation is still τ.

An alternative way to see this is to extend to population inference. Let $(\mu_{1j}, \sigma_{1j}^2)$ be the population mean and variance, respectively, of $Y_i(1)$ for all cases in subclass

j. Let $(\mu_{0j}, \sigma^2_{0j})$ be similarly defined population quantities for all $Y_i(0)$ in subclass j. If we think of $y^*_i = Y_{i'}(w_i)$ as randomly drawn from the correct populations, the expectations of $\bar{y}^*_{1j}$ and $\bar{y}^*_{0j}$ continue to be μ_{1j} and μ_{0j}, respectively.

2.3.2 Matching subclass and non-matching treatment status

The record linkage algorithm might link two records with similar covariates, and hence the same subclass, that receive different treatments. This can create substantial problems for causal estimates. To illustrate, suppose instead that David receives the treatment but Davy does not. We then attribute Davy's outcome to receiving the treatment rather than not receiving it. This incorrect link biases the estimated treatment effect in the opposite direction of τ. In fact, if we make this mistake many times, we could end up concluding that the treatment has the opposite effect than it truly does.

To demonstrate this, suppose that we erroneously set a single treated individual's $y^*_i = Y_{k'}(0)$ where $(i, k') \in \mathcal{S}_j$. We then have

$$\begin{aligned} E(\hat{\tau}^*_j) &= \sum_{i \in \mathcal{S}_j, i \neq k'} \left(\frac{n_{1j}/n_j}{n_{1j}} Y_i(1) - \frac{1 - n_{1j}/n_j}{n_{0j}} Y_i(0) \right) + \frac{n_{1j}/n_j}{n_{1j}} Y_{k'}(0) - \frac{1 - n_{1j}/n_j}{n_{0j}} Y_{k'}(1) \\ &= (1/n_j) \left(\sum_{i \in \mathcal{S}_j, i \neq k'} \tau_i - \tau_{k'} \right). \end{aligned} \tag{2.12}$$

In terms of population quantities, the contribution of record i to $E(\bar{y}^*_1)$ is $w_i \mu_{0j}/n_{1j}$ rather than $w_i \mu_{1j}/n_{1j}$, and the contribution to $E(\bar{y}^*_0)$ is $(1 - w_i)\mu_{1j}/n_{0j}$ rather than $(1 - w_i)\mu_{0j}/n_{0j}$. The result is that $\hat{\tau}^*$ is biased.

2.3.3 Non-matching subclass and matching treatment status

We next consider when the linking error impacts the subclass assignment but not the treatment assignment. Suppose that we link record i in File A named Anna Karenina with record i' in File B named Alexis Karenin. Anna is 30 years old, female and

has 2 children; Alexis is 40 years old, male and has 1 child. Neither received the treatment. When we incorrectly link Anna with Alexis, we put Alexis's $y_{i'}$ with Anna's (x_i, w_i). Hence, when the propensity score model is reasonable, $y_{i'}$ could be placed in an incorrect subclass, but with the correct treatment status (in this case, control). This adds bias to the treatment effect estimate.

It is difficult to characterize the nature of this bias, since it depends on how similar the covariate distributions in the incorrectly matched subclass are to those in the actual subclass. What is clear is that it no longer makes sense to consider treatment assignment as completely random within the subclasses, since it is no longer reasonable to believe that covariates are balanced. Hence, it is cumbersome to derive mathematical arguments like those in Chapters 2.3.1 and 2.3.2. However, we can gain some insight into this bias when we suppose that the process generating the linkage errors is independent of the values of x and $(Y(1), Y(0))$. In this case, we can consider the erroneous link for record i to be selected randomly from cases with matching w in the incorrect subclasses. Using the population quantities and averaging over subclasses, the contribution of record i to $E(\bar{y}_1^*)$ is $w_i \mu_{1(-j)}/n_{1j}$, where $\mu_{1(-j)} = \sum_{h \neq j} \mu_{1h}(n_{1h}/(n_1 - n_{1j}))$, rather than μ_{1j}/n_{1j}. Similarly, the contribution to $E(\bar{y}_0^*)$ is $(1 - w_i)\mu_{0(-j)}/n_{0j}$ where $\mu_{0(-j)} = \sum_{h \neq j} \mu_{0h}(n_{0h}/(n_0 - n_{0j}))$ rather than μ_{0j}/n_{0j}. Indeed, in an extreme case, if all records are subject to this error then one might as well not even have used subclassification, in which case $\hat{\tau}^* \approx \mu_1 - \mu_0$, where μ_1 and μ_0 are the marginal population averages of the treated and control outcomes.

2.3.4 Non-matching subclass and non-matching treatment status

Finally, we consider the case of wrong treatment and wrong subclass. To illustrate, we link the record of an Adam Trask in file A with an Aron Trask in file B. Adam is 60 with two children, and Aron is 18 with no children. Adam has received the blood pressure medication but Aron has not. When we link Aron's outcome with Adam's

background covariates and treatment indicator, we observe an incorrect link of outcome and subclass as well as an incorrect link of outcome and treatment indicator.

As in Chapter 2.3.3, the potential bias induced by this type of linkage error is difficult to characterize. When linkage errors are independent of x and $(Y(1), Y(0))$, we can use arguments like those in Chapter 2.3.3. Averaging over subclasses, the contributions of record i to $E(\bar{y}_1^*)$ and $E(\bar{y}_0^*)$ are $w_i\mu_{0(-j)}/n_{1j}$ and $(1-w_i)\mu_{1(-j)}/n_{0j}$, respectively. If we make this type of mistake many times, the result will be as if we never subclassified, and we additionally labeled the treated group as the control group and vice versa, resulting in $\hat{\tau}^* \approx \mu_0 - \mu_1$.

2.4 Case selection rules for threshold linkage

Clearly, linkage errors can have negative consequences for causal inference. One could restrict causal inference to estimating only with cases known to be true links with complete certainty. However, this may exclude some links that are correct, or possibly innocuously in error like those in Chapter 2.3.1, ultimately inflating mean squared errors. We thus need a rule for deciding which linked pairs to use in $\hat{\tau}^*$.

One approach is to attempt to choose the threshold for accepting links to minimize the mean squared error of $\hat{\tau}^*$. If we add links sequentially in decreasing order of their linkage scores, we would expect that adding the first few records to the known links should result in adding a sizable proportion of correct links. The mean squared error of $\hat{\tau}^*$ should decrease as we add cases to (2.9) until we start adding many non-links, when bias introduced by the invalid links can overwhelm the reductions in variance due to increased sample size.

Unfortunately, an estimator for the mean squared error of $\hat{\tau}^*$ is not apparent, as we do not know the value of τ. Instead, we turn to a quantity that has similar behavior as the mean squared error, is easy to compute, and is familiar to users of propensity score subclassification: the estimated variance in (2.10). In particular,

we present three algorithms for selecting cases based on estimated variances. As a reminder, in addition to SUTVA and strong ignorability, we derive the algorithms under the assumptions that (i) linkage errors are independent of x and $(Y(1), Y(0))$, (ii) propensity score subclassification results in groups with balanced covariate distributions, and (iii) treatment effects are additive. We assume that the analyst uses a threshold based record linkage technique like those described in Chapter 2.2.2.

2.4.1 The minimum estimated variance algorithm

The first algorithm, which we call the minimum estimated variance or MEV algorithm, is initialized as follows. For each record pair (i, i') we compute $S(\gamma_{ii'})$ using (2.7), identifying the highest scoring link for each. When the same record from File B is the top link for multiple records in File A, one can use a post-processing strategy that enforces one-to-one linkage (Herzog et al., 2007). For computational convenience, in the simulation studies we allow records in File B to be selected as the top link multiple times if necessary; this has very minimal impact on results as we describe in Section 2.5. Let $\mathcal{L}_0$ be the set of the l_0 record pairs known with certainty to be correct links. We then compute $\hat{\tau}^*$ using (2.9) and the estimated variance using (2.10), calculating the λ weights and other statistics in (2.9) and (2.10) from the cases in $\mathcal{L}_0$. We use the propensity scores determined from the analysis of all of File A.

We next arrange the top pairs in descending order of $S(\gamma_{ii'})$. Let $[h]$ index the rank order of the hth record pair, so that [1] is the pair not in $\mathcal{L}_0$ with the highest linking score, [2] is the pair not in $\mathcal{L}_0$ with the second highest linking score, and so on. Set a counter $h = 1$. We append record pair $[h]$ to $\mathcal{L}_{h-1}$ to create $\mathcal{L}_h = \mathcal{L}_{h-1} \cup (x_{[h]}, w_{[h]}, y^*_{[h]})$. We repeat this process for all pairs with linking scores above some minimum threshold, as values below this threshold are considered known not to be links, each time incrementing h by one. As a result, we have a collection

of $L \leqslant (n_A - l_0)$ successively larger sets $\mathcal{L}_h$. We evaluate (2.9) and (2.10) for each set of cases in $\mathcal{L}_h$, re-computing λ each time but using the propensity scores based on all of File A. We select $\mathcal{L}_{min} = \{\mathcal{L}_h : h = \arg\min_h \hat{var}(\hat{\tau}^*)\}$.

As we add correct links, the estimated variance tends to decrease due to the increase in sample size. In fact, even adding errors like those in Chapter 2.3.1 still can result in decreased estimated variance, as we generally add sample size while still drawing from the correct marginal distributions of the outcomes within each subclass. However, when we add pairs with other types of linkage errors, we add draws from incorrect marginal distributions, causing the estimated variance to tend to increase. Thus, the MEV procedure tends to favor adding cases that are correct links and links with errors like those in Chapter 2.3.1 and to disfavor adding incorrect links of other types.

To gain further insight, we present a rough approximation to the expected value of (2.10). For simplicity, we ignore uncertainty due to estimating propensity scores and subclass boundaries, and treat observations within subclasses as independent. To motivate why the criterion is useful, suppose we consider the linked data in $\mathcal{L}_h$ as a sample from a hypothetical population of linked datasets using that threshold. For $j = 1, \ldots, J$, let $S^{2*}_{h0j} = E(s^{2*}_{h0j})$ and $S^{2*}_{h1j} = E(s^{2*}_{h1j})$, where we add the subscript h to emphasize that the quantities are computed with the cases in $\mathcal{L}_h$. For any $\mathcal{L}_h$, let $\mathcal{C}_{hj}$ be the set of correct links and p_{hj} be the probability of a randomly sampled link within subclass j being correct. Within any subclass j, we can write S^{2*}_{h0j} with an iterated variance,

$$S^{2*}_{h0j} = E(Var(y^* \mid w = 0, \mathcal{C}_{hj})) + Var(E(y^* \mid w = 0, \mathcal{C}_{hj})). \tag{2.13}$$

Let μ_{h0j} and σ^2_{h0j} be the population mean and variance of y^*_i for all erroneously linked records with $w_i = 0$ when using the threshold associated with $\mathcal{L}_h$. For the first term

of (2.13), we have

$$E(Var(y^* \mid w = 0, \mathcal{C}_{hj})) = S^2_{h0j} p_{hj} + \sigma^2_{h0j}(1 - p_{hj}), \tag{2.14}$$

where S^2_{h0j} is the variance of y for records in subclass j that are true links with $w = 0$. For the second term of (2.13), we have

$$Var(E(y^* \mid w = 0, \mathcal{C}_{hj})) = (\mu_{0j} - \mu_{h0j})^2 p_{hj}(1 - p_{hj}). \tag{2.15}$$

We can derive a similar expression for S^{2*}_{h1j}.

Putting it all together, we have the approximation,

$$\begin{aligned} E(\hat{var}(\hat{\tau}^*_j)) \;\approx\; & p_{hj}\left(\frac{S^2_{0j}}{n_{h0j}} + \frac{S^2_{1j}}{n_{1hj}}\right) + (1 - p_{hj})\left(\frac{\sigma^2_{h0j}}{n_{h0j}} + \frac{\sigma^2_{h1j}}{n_{h1j}}\right) \\ + \;& \left(\left(\frac{1}{n_{h0j}}(\mu_{0j} - \mu_{h0j})\right)^2 + \left(\frac{1}{n_{h1j}}(\mu_{1j} - \mu_{h1j})\right)^2\right) p_{hj}(1 - p_{hj}) \end{aligned} \tag{2.16}$$

The first term in (2.16) is the variance for correct links within the subclass, weighted by the proportion of correct links. The second term is a variance contribution from the incorrect links. Generally, we expect the σ^2_{hwj} to exceed the corresponding S^2_{wj}, since for any (w, j) the distribution of y^*_i for incorrect links generally should be more dispersed than the corresponding distribution of y_i for correct links, as evident from the consequences of linkage error described in Chapter 2.3. The third term can be viewed as a penalty for introducing incorrect links.

Using (2.16), we see that (2.10) tends to be smallest in expectation when p_{hj} is large and when linkage errors do not cause substantial differences between the means and variances of the outcomes for the correct and incorrect link cases. Thus, using the criterion should favor thresholds where the fraction of true links is high and the consequences of mistakes are low, which can help improve the accuracy of treatment effect estimates compared to using only cases known to be true links.

2.4.2 The estimate-tethered stopping rule algorithm

While MEV penalizes bias as desired, it has the potential to result in undesirable case selection decisions. To see this, consider a scenario where the number of correct links is small compared to the number of incorrect links, and the treatment effect is small relative to the marginal variance of the outcome variable. In this case, (2.16) could be smallest when one includes as many links as possible. Adding cases, even incorrect links, increases the sample sizes used in (2.16), which could reduce the variance terms in (2.16) by enough to overwhelm the increase caused by bias.

To reduce the potential for such undesirable selections, we restrict $\mathcal{L}_h$ to sets where the corresponding $\hat{\tau}^*$ is within k standard errors of the estimated treatment effect based on links known to be correct. From this set, we select the $\mathcal{L}_h$ with the minimum estimated variance. We call this the estimate-tetheredstoping rule, abbreviated as ETSR. Formally, we choose $\mathcal{L}_{ETSR} = \{\mathcal{L}_h : h = \arg\min_h \hat{var}(\hat{\tau}^*),\ \hat{\tau}^*_{\mathcal{L}_h} \in \hat{\tau}^*_{\mathcal{L}_0} \pm k\sqrt{\hat{var}(\hat{\tau}^*_{\mathcal{L}_0})}\}$. We use $k = 0.5$ in the simulation results reported in Chapter 2.5, but the results are relatively robust to choices of k between 0.5 and 2. We present results for ETSR under different values of k in the supplement. Generally we expect ETSR to result in more conservative linkage than MEV, but with less room for bias.

2.4.3 The minimum estimated difference-in-outcomes variance algorithm

The MEV and ETSR use the propensity score subclassification when computing their respective criteria. As suggested by reviewers, some analysts may prefer to separate the propensity score analysis from the linkage decisions as much as possible; see Chapter 2.6 for additional discussion of this point. We therefore propose the minimum estimated difference-in-outcomes variance algorithm, which we abbreviate MEDOV. We select $\mathcal{L}_{MEDOV} = \{\mathcal{L}_h : h = \arg\min_h \hat{var}(\bar{y}^*_{1h} - \bar{y}^*_{0h})\}$, where $\bar{y}^*_{1h}$ and $\bar{y}^*_{0h}$ are the marginal means of the treated and control units in $\mathcal{L}_h$. MEDOV is similar

to MEV, but we estimate the variance before subclassification, i.e., use only $J = 1$ class in (2.10).

2.5 Simulation studies

Here we present results of simulation studies evaluating the performance of the case selection algorithms from Chapter 2.4. We base all simulations on the RL10000 data from the "RecordLinkage" package in R (Borg and Sariyar, 2015). This dataset includes full name separated into four fields and birth dates on 9000 individuals. For 1000 of these 9000 individuals, the RL10000 dataset also includes duplicate records with typographical errors on some of the fields. No other variables are available on the file. In all simulations, we use first name, last name, birth month, and birth day as linking variables.

In the simulations reported here, we split the 10000 records into File A comprising $n_A = 2000$ records and File B comprising $n_B = 8000$ records. In each simulation run, File A includes the 1000 records with duplicates and a random sample of 1000 records without duplicates. File B includes the 1000 duplicates with errors and the remaining 7000 records without duplicates. Due to the random sampling of non-duplicates across runs, the threshold values and linkage quality can change across the simulations. The effects of such changes are minor.

For each of the 1000 records in File A with duplicates in File B, we modify the birth year of its true link in File B so that both have the identical birth year. This allows us to block on birth year, i.e., require pairs to have the same birth year if they are to be considered links, and link on first name, last name, birth month, and birth day. Blocking on birth year reduces the comparison space, which improves the quality of links and reduces computational time. Blocking is a standard practice in record linkage settings (Herzog et al., 2007). We also allow units in File B to be linked to more than one unit in File A, primarily for computational convenience in

repeated simulation studies. This has minimal impact on the simulation results, as typically only zero to two duplicates are used for the thresholds with high (greater than 95%) link rates.

In all simulations, we present results based on $\hat{\tau}^*$ and its variance without any regression adjustments; results for regression adjustment are in the supplementary material. We estimate propensity scores using a logistic regression with treatment indicator as the outcome and main effects of the covariates as predictors. We use the full data set in File A to calculate the propensity scores and $\mathcal{L}_{h0}$ to calculate the subclass boundaries.

In any $\mathcal{L}_h$, we use the sample sizes in the linked data to calculate each λ_j. Codes for the simulation are available online at https://github.com/jodywortman/simultaneous-record-linkage-and-causal-inference.

One could instead estimate the subclass boundaries, or possibly even the propensity scores themselves, using only the data in $\mathcal{L}_h$. This potentially could improve the covariate balance within the subclasses for that $\mathcal{L}_h$. We do not do so here, as manual adjustment of the quantiles is not practical in repeated simulation studies. Nonetheless, we recommend that analysts check covariate balance in any application, so as to avoid poorly balanced subclasses. Incorporating such checks into the algorithms for $\mathcal{L}_h$ is an intriguing topic for future research.

As a baseline, we compare results to treatment effect estimates that would be obtained if all cases in File A were perfectly linked, i.e., all 1000 records with a link are put together and the remainder are designated non-links. We refer to these as true links or correct links, and refer to the results as the "Perfect" results. We also compare results to the most conservative linkage strategy, in which we use only those record pairs for which all fields agree perfectly. We call these as known links or exact links, and refer to results as the "Known" results.

2.5.1 Data generation and linkage methods

As RL10000 has no other variables, for each record i, we generate its treatment indicator w_i, two covariates (x_{i1}, x_{i2}), and outcome y_i as follows. We sample each w_i from a Bernoulli distribution with probability .5. Given w_i, we sample x_{1i} from a Poisson distribution with mean $(8-3*w_i)$. We sample x_{2i} from a normal distribution with mean $-w$ and standard deviation 3. In this way, the distribution of x_i differs for treated and control units, making propensity score subclassification useful compared to estimating τ as the difference in the marginal means.

We generate outcomes according to six different scenarios, each with an additive treatment effect. In the first four, we use a linear response surface

$$y_i = 5 + 5x_{i1} + 3x_{i2} + \tau w_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2). \tag{2.17}$$

In the first scenario, we set $(\tau = 50, \sigma = 10)$ to represent a treatment effect that is large relative to the variance of the outcome. In the second scenario, we set $(\tau = 10, \sigma = 10)$ to assess the impact of having a more modest treatment effect. In the third scenario, we set $(\tau = 1, \sigma = 10)$ to examine the impact of a small, but still non-zero, treatment effect relative to the variance of the outcome. In the fourth scenario, we set $(\tau = 50, \sigma = 25)$ to assess the impact of increasing the variance of the outcome when the treatment effect is large. In the fifth scenario, we make the covariates have a stronger association with the outcome, using

$$y_i = 5 + 15x_{i1} - 7x_{i2} + \tau w_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2) \tag{2.18}$$

with $(\tau = 50, \sigma = 10)$. Finally, in the sixth scenario, we assess the impact of having a non-linear response surface, using

$$y_i = 5 + 0.2x_{i1}^2 + exp(0.7x_{i2}) + \tau w_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2) \tag{2.19}$$

with $(\tau = 50, \sigma = 10)$.

For each simulation run, we resample values of $(w_i, x_{1i}, x_{2i}, y_i)$ for all records. In all scenarios, using propensity score subclassification improves the causal estimates substantially. For example, in the first scenario where $(\tau = 50, \sigma = 10)$ and the response surface is linear, the difference in marginal means of the outcome for treated and control cases is around 31. Using the true links for all cases in File A, $\hat{\tau}$ is around 47. Thus, subclassification allows for substantial bias reduction.

To implement the Fellegi-Sunter record linkage, we first block on birth year, requiring links to have the same values of birth year. We then compare the first and last names of pairs of records using Jaro-Winkler scores, which are dichotomized and fed into the Fellegi-Sunter algorithm. For each name $f \in \{1, 2\}$, we classify $\gamma_{fii'} = 1$ when $\Phi_{JW}(fii') > .95$ and $\gamma_{fii'} = 0$ otherwise; that is, the Jaro-Winkler score must exceed .95 for the fields to be called in agreement. We compare birth month and birth day using binary exact agreement indicators. We compute linkage scores for each record pair that agrees on birth year using (2.7). We set $\theta_{mf} = .95$ for all f and set θ_{uf} as frequency of agreement in field f. We consider record pairs with $S(\gamma_{ii'}) < 0$ in (2.7) not to be links.

2.5.2 Results

Table 2.1 summarizes the quality of links at different thresholds for the first simulation scenario. Results are similar for other scenarios. Link rate corresponds to the percentage of links that correctly correspond to the same person. Units refers to the number of cases in the linked dataset, and duplicates refers to the number of non-unique appearances of a person from File B. Linkage quality at thresholds of 9.3 and above is high but deteriorates quickly as one drops the threshold, with more false links and duplicates. The supplementary material includes examples of the linked data under different thresholds, illustrating the types of errors tolerated when decreasing the threshold. Of course, in applications we generally are not able to de-

Table 2.1: Linkage summary under various thresholds for first simulation scenario.

Threshold	Link Rate	Units	Duplicates
0.3	76.2 %	1309	27
0.8	84.6 %	1179	8
1.6	91.6 %	1088	3
2.1	95 %	1048	0
2.8	97.5 %	1018	0
9.3	98.7 %	1002	0
9.8	99.2 %	973	0
10.5	99.7 %	775	0
11.8	100 %	638	0
19.5	100 %	531	0

termine the link rates at different thresholds, and hence not able to identify optimal threshold values. This motivates consideration of the case selection procedures.

Figure 2.2 summarizes the distributions of $\hat{\tau}^*$ and its estimated variance based on (2.10) for 100 independent runs of the first simulation with a linear response surface and $(\tau = 50, \sigma = 10)$ at all qualifying values of the threshold for selecting cases. The horizontal line in top panel corresponds to $\hat{\tau}$ with the true links. The choice of threshold matters for the quality of the causal estimate. Using thresholds below 9.3 includes incorrect links that degrade the accuracy of $\hat{\tau}^*$. On the other hand, using the highest threshold values cause $\hat{\tau}^*$ to be based on relatively small numbers of individuals, which results in the largest variances of $\hat{\tau}^*$. Apparently, the sweet spot reflecting a close-to-optimal trade off in contributions to mean-squared error is a threshold somewhere around 9.8, which provides point estimates clustered most closely around the value of $\hat{\tau}$ attainable with the true links. As is evident in the bottom panel of Figure 2.2, across the 100 runs the value of (2.10) tends to be minimized when the threshold is around 9.8, suggesting that using the case selection algorithms could reduce mean squared errors.

Table 2.2 displays key results of the simulation runs. $\text{Var}(\hat{\tau}^*)$ refers to the empirical variance of the estimated treatment effects across each set of 100 runs, and

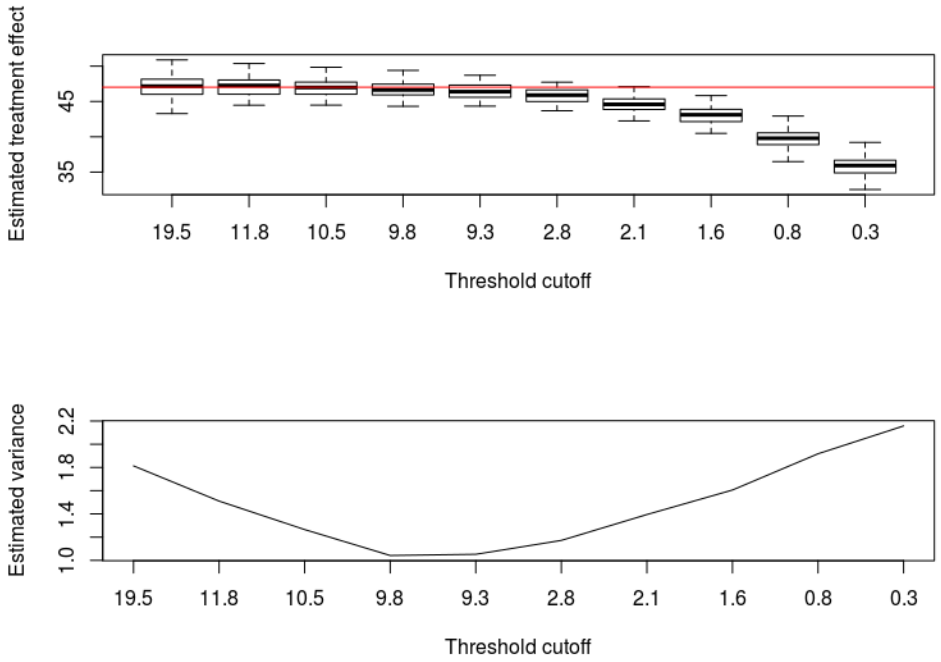

FIGURE 2.2: Distribution of 100 point and variance estimates of treatment effects for simulation with Fellegi-Sunter linkage and constant treatment effect, where $\tau = 50$ and $\sigma = 10$.

$\hat{var}(\hat{\tau}^*)$ refers to the average of the estimated variances based on (2.10) across each set of 100 runs. Threshold refers to the average value of the threshold chosen across the 100 runs. "Perfect" refers to the analysis using all 1000 true links and "Known" refers to the analysis with all records in File A with perfect agreement among all linking variables. Turning first to the simulation setting with $(\tau = 50, \sigma = 10)$, all three case selection methods reduce mean squared errors compared to using only known links. All of the case selection methods have increased mean squared errors compared to using the true links, reflecting the information loss from having to use inexact linkage. The percentage increase in mean squared error ranges from 11% to 28% for the case selection methods, whereas it is around 56% for using only the known links. ETSR offers the most substantial reductions in mean squared error, although all three methods have comparable performances. The average thresholds

selected by MEV and ETSR are around 9.1 and 10.5, respectively. MEDOV is more conservative, in that it adds the smallest number of links to the known links.

Turning to the second and third scenarios where we reduce τ, the three case selection methods offer substantial reductions in mean squared errors compared to using only the known links. The reductions in mean squared error for MEDOV are not as substantial as those for MEV and ETSR, mostly because it does not add many links to the known cases as evident in the large average thresholds. In the scenario with $\tau = 1$, interestingly, MEV has a smaller mean squared error than using all the true links. This results partly because τ is close to zero. In this scenario, MEV ends up using some incorrect links that bias the treatment effect toward zero, which is close enough to τ to reduce mean squared error compared to using all the true links.

In the fourth scenario where we return $\tau = 50$ and increase the variance to $\sigma = 25$, we again see that the case selection methods have larger mean squared errors than using the true links, as one would expect generally. Here, however, the MEV has a somewhat larger bias, which happens because the MEV criterion accepts too many false links, as evident by the average threshold level of 6.7. Because of the large variance in the outcomes, accepting false links tends to have greater impact on the bias and mean squared error in $\hat{\tau}^*$. This illustrates the concerns about MEV noted at the beginning of Chapter 2.4.2 and used to motivate ETSR. In contrast, ETSR and MEDOV continue to substantially outperform using the known links only, with ETSR offering slightly larger reductions than MEDOV.

In the fifth scenario, the stronger associations between the covariates and the outcome increase variances, and hence mean squared errors, of the treatment effect estimators compared to the previous scenarios. Here, ETSR has the smallest mean squared error among the case selection procedures. In contrast, MEDOV performs worse than using the known cases alone.

In the sixth scenario where the response surface is non-linear, all three case se-

lection procedures are again preferable to using known links alone. Again, ETSR offers the greatest reductions in mean squared error, getting to almost the same mean squared error as using the true links. Evidently, tethering the estimates helps ensure that low quality links are not added to the sample used for the treatment effect estimate.

Looking across all six scenarios in Table 2.2, it is apparent that (2.10) tends to underestimate the theoretical variance of $\hat{\tau}^*$. This is not surprising for multiple reasons. First, (2.10) does not properly account for uncertainty in the linkage. This is typical of analyses based on linked data, which rarely account for this source of uncertainty. Second, (2.10) inherits the biases in (2.2) with perfectly linked data. Third, the formula in (2.10) does not explicitly account for the effects of the case selection. Developing a reliable estimate of the theoretical variance of $\hat{\tau}^*$ is an important area for future research.

The supplement contains results of additional simulations where the core assumptions are satisfied, including a scenario where File B comprises $n_B = 2000$ records and a scenario where we use Jaro-Winkler linkage. Results are qualitatively similar to those presented here. It also includes three scenarios where some assumptions are violated. Specifically, we run simulations where the treatment effect is not constant but linkage errors are still independent of all variables, where the treatment effect is not constant and linkage errors are correlated with a causally relevant variable, and where one of the linking variables is a confounder in the causal analysis. In all three scenarios, the case selection methods offer smaller or approximately the same mean squared errors compared to using only the known cases.

2.6 Conclusions

Methods for causal inference and record linkage have developed independently, but the simulation results indicate that it can be fruitful to consider methods that explic-

Table 2.2: Summary of results across 100 runs of the six simulation scenarios with Fellegi-Sunter linkage. Results with the lowest MSE for each simulation are in bold.

Scenario	Linkage	Mean $\hat{\tau}^*$	Var ($\hat{\tau}^*$)	$\hat{var}(\hat{\tau}^*)$	MSE $\hat{\tau}^*$	Threshold
Linear, ($\tau = 50, \sigma = 10$)	Perfect	47.0	1.0	0.9	9.9	
	Known	47.0	6.8	6.9	15.4	
	MEV	46.6	1.1	1.0	12.7	9.1
	ETSR	**47.0**	**1.9**	**1.2**	**11.0**	**10.5**
	MEDOV	46.9	1.7	1.4	11.4	13.1
Linear, ($\tau = 10, \sigma = 10$)	Perfect	7.1	1.1	0.9	9.3	
	Known	7.3	7.2	6.5	14.3	
	MEV	7.0	1.1	1.0	9.9	7.5
	ETSR	**7.1**	**1.7**	**1.1**	**9.9**	**8.2**
	MEDOV	6.8	1.7	1.3	12.1	11.1
Linear, ($\tau = 1, \sigma = 10$)	Perfect	-1.9	1.1	0.9	9.3	
	Known	-1.7	7.2	6.5	14.3	
	MEV	**-1.8**	**1.1**	**1.0**	**8.9**	**7.6**
	ETSR	-1.8	1.7	1.2	9.8	9.1
	MEDOV	-2.1	1.6	1.4	10.9	12.7
Linear, ($\tau = 50, \sigma = 25$)	Perfect	47.3	4.8	4.0	12.1	
	Known	47.7	33.7	27.4	38.7	
	MEV	46.0	10.5	4.1	26.6	6.7
	ETSR	**47.0**	**8.1**	**4.8**	**17.1**	**8.8**
	MEDOV	46.9	7.4	5.9	17.2	13.8
Linear, High R^2, ($\tau = 50, \sigma = 10$)	Perfect	43.9	6.5	7.0	43.7	
	Known	44.5	48.9	46.9	79.0	
	MEV	42.5	13.4	7.1	69.3	6.1
	ETSR	**43.8**	**12.9**	**8.1**	**51.5**	**7.9**
	MEDOV	41.7	21.6	9.5	90.8	9.8
Non-linear, ($\tau = 50, \sigma = 10$)	Perfect	48.1	22.7	25.8	26.3	
	Known	48.1	96.3	104.7	99.0	
	MEV	43.3	29.2	5.8	74.4	5.8
	ETSR	**47.0**	**17.8**	**9.5**	**26.7**	**9.8**
	MEDOV	45.6	29.6	7.2	48.8	9.8

itly account for both tasks. For settings where covariates and assignments are in one file and outcomes are in the other file, the simulations here and in the supplementary material suggest that case selection strategies can improve causal estimates for analyses based on propensity score subclassification. In these simulations, arguably ETSR performs best overall. It has the smallest mean squared errors in some scenarios, and when other case selection rules have lower mean squared errors, ETSR generally is not far behind. Of course, as with any simulation study, these findings are based on

limited simulation studies and particular assumptions. Additional research is needed to assess the performance of the case selection strategies when these assumptions are not reasonable.

As with any methodology, there are scenarios where the case selection procedures may not be effective. In particular, when the known links include outliers that pull the estimate of $\hat{\tau}_{\mathcal{L}_0}$ far away from τ, the procedures might not add many cases, even true links, to the data used in the causal estimate, as doing so could cause the estimated variance to increase. Additionally, the case selection procedures may give misleading results when their underlying assumptions, in particular independence of the linking variables and constant treatment effects, are unreasonable. Finally, the procedures may suffer when the underlying analysis models are poorly specified, including the propensity score models, subclass boundaries, and regressions for adjusted inferences.

The case selection procedures are designed to work with common record linkage techniques like the Fellegi-Sunter approach. A potential alternative is to adapt record linkage techniques that average over different compositions of the linked population. For example, Gutman et al. (2013) and Dalzell and Reiter (2018) sample from the posterior distribution of a latent linking matrix, informed by a posited regression model that connects an outcome variable in File A to predictors in File B. Adapting such approaches specifically for causal inference is an intriguing area for future research.

We recommend being sensible in the choice of linkage technology. We found that adding many incorrect links, e.g., by using very low thresholds to accept almost any proposed link, can reduce the estimated variance of the treatment effect due to the increased sample size. However, this results in poor quality estimates of τ. When reasonable cutoffs on linkage scores are enforced, the pattern of the estimated variance under varying thresholds tends to be U-shaped. However, the estimated

variance can become S-shaped when large numbers of incorrect links are added. Using the ETSR limits the possibility of favoring thresholds corresponding to high numbers of incorrect links, but we still emphasize the importance of using sound record linkage techniques when making causal inferences with linked data.

Finally, we close with a comment on the philosophy of causal inference in observational studies. Many researchers follow the guidance to separate the design of the study from the analysis (Imbens and Rubin, 2015). The case selection procedures partially adhere to that guidance. When the covariates and treatment are in the same file, one can estimate propensity scores and form subclasses without referring to the outcomes. However, the procedures utilize the outcomes when selecting the sample to use for estimation. If one seeks the potential gains in accuracy from adding more links, this is the price to pay for working with imperfect data.

3

BRACS: Model and Simulations

3.1 Introduction

Researchers and analysts often seek to combine data sources to draw inferences. For example, one may seek to merge student lists and voter registration records to study political engagement among college students, merge health records from two different providers to see a patient's medical history, or merge enrollment lists for social service programs to maximize program enrollment. The data sources may be from planned studies (e.g., from a medical trial or a controlled experiment), or data collected for a different purpose, such as National Change of Address (NCOA) data. Ideally, one can use unique identifiers, such as social security numbers, for linking, but in many cases this information is either not available or unreliable. Researchers then have to use imprecise information such as names, birth dates, or demographic descriptors to merge the datasets.

A common framework for probabilistic record linkage was introduced by Fellegi and Sunter (1969). They use a mixture distribution to estimate linkage probabilities. As an example, imagine linking across two datasets, A and B, both of which have

birth date, city of birth, and city of current residence. The model of Fellegi and Sunter (1969) assumes that for each possible pairing of records (A_i, B_j), the agreement between shared fields comes from two distributions–one if A_i and B_j belong to the same individual, and another if they do not. Complications arise when incorporating string data (Newcombe et al., 1959; Jaro, 1989, 1995; Larsen and Rubin, 2001). One solution is to keep the categorical comparison structure, but to use a similarity metric such as normalized Levenshtein similarity split into intervals (Levenshtein, 1966; Winkler, 1990; Sadinle, 2017). Sadinle and Fienberg (2013) and Sadinle (2017) build on the Fellegi-Sunter framework, generalizing it to a Bayesian record linkage approach that incorporates string data and allows for different levels of comparison across fields.

These approaches assume one common distribution given the link status. However, one easily can imagine scenarios where this may not hold. For example, younger people may move more frequently than older people, resulting in a lower rate of address agreement even among true links. Typical procedures would not account for these nuances in the linkage.

In this chapter, we introduce a Bayesian model for record linkage designed for data in which there are differences in the comparison distributions for links and non-links, conditional on attributes observed in one file. We call this Bayesian Record linkage and Clustered Sub-models (BRACS). The chapter is structured as follows. In Chapter 3.2, we introduce notation and motivate modification of existing Bayesian record linkage approaches. In Chapter 3.3, we introduce BRACS as a model for the linkage and outline steps for a Gibbs sampler to estimate the model parameters. In Chapters 3.4 and 3.5, we use simulated data to evaluate how BRACS compares to Bayesian record linkage without a clustered structure. In Chapter 3.6, we discuss the results of the simulations and the benefits and limitations of the method.

This work was motivated by an application, which we will discuss in more depth

in Chapter 4, of linking records from North Carolina's list of registered voters (the voter file) to a list of voters who cast provisional ballots in the 2016 general election. The two data sets each have first, last, and middle name, address, political party affiliation, gender, and race, and the voter file alone has age. We use this example throughout this chapter to explain aspects of the BRACS model.

3.2 Notation

Suppose we have two datasets, A and B, with N_a and N_b records, respectively. A is comprised of records A_i, where $i \in (1, ..., N_A)$, and B is comprised of records B_j, where $j \in (1, ..., N_B)$. Let A_{i0} be a latent identifier for record A_i. Similarly, let B_{j0} be a latent identifier for record B_j. When records A_i and B_j belong to the same individual or entity, so that $A_{i0} = B_{j0}$, we say that they are a *link*.

One way to represent which record pairs are links is in matrix form. Let Δ be a matrix of size $N_A \times N_B$ with entries $\Delta_{ij} = 1$ when $A_{i0} = B_{j0}$ and $\Delta_{ij} = 0$ otherwise. If we are linking A to a larger B and assume that there is only one possible link to each A_i in B, we can map Δ to a vector Z of length N_A, where $Z_i = j$ when (A_i, B_j) are linked. We define $Z_i = N_B + 1$ when A_i does not have a corresponding record in B.

We suppose that A and B have F fields in common, e.g., demographic information, name, or address; these are also called linking variables. In order to determine whether A_i and B_j belong to the same individual, we measure similarity between these linking variables. For $f = 1, ..., F$, let $D_f(i, j)$ be the similarity between A_i and B_j, with f indexing the field. There may be multiple possible values of $D_f(i, j)$, depending on the field and comparison type. For example, we may consider a categorical f such as state of residence and look at binary agreement only (e.g., is OHIO the same state as NORTH CAROLINA, or is it different?). When f is a name field, however, we may calculate a string similarity (e.g., how similar is the name

ZACHARY to ZACK?) and split it into L_f possible categories of $D_f(i,j)$. Thus we may differentiate, for example, between the same name, e.g., ZACK and ZACK, or very similar names (potentially different spellings like ZAK or ZACH), or completely different names (ZACK and JEFFERY). For each record pair and each f, we say that $\gamma_{ij}^f = l$ if $D_f(i,j)$ fits into the l bucket of agreement. In total, these comparison indicators comprise $\gamma_{ij} = (\gamma_{ij}^1, \gamma_{ij}^2, ...\gamma_{ij}^F)$.

We assume that each γ_{ij} is a random realization from a distribution Γ_{ij}. One modeling approach is to use a mixture model for Γ_{ij}. Building off of notation from Sadinle and Fienberg (2013), in this approach we assume that

$$\Gamma_{ij}|\Delta_{ij} = 1 \sim iid\ G_1 \tag{3.1}$$

$$\Gamma_{ij}|\Delta_{ij} = 0 \sim iid\ G_0. \tag{3.2}$$

When (A_i, B_j) is a link, the field comparisons are iid draws from some distribution G_1 for all linked (i,j). When (A_i, B_j) is not a link, the field comparisons are iid draws from some distribution G_0 for all non-linked (i,j).

We now extend this framework to scenarios in which there are linkage clusters within the pairs, each with its own distribution. For example, one linkage cluster might comprise younger people and another older people. Let K be the number of linkage clusters. In our application, whether or not each pair belongs to a given cluster $k \in (1, ..., K)$ is uniquely determined by values of fields in B. Thus, the cluster is known for each pair (i,j) as values in record B_j imply the cluster k. We therefore can split records in B into K disjoint sets, such that $B = \bigcup_{k=1}^{K} S_k$, where S_k is the set of indexes j of records that fall in cluster k. For example, suppose we believe comparison distributions depend on age, which is unique to B. Suppose records $(1, 2, 30, 190)$ in B all belong to young people, implying category $k = 1$, then $j = 2 \implies j \in S_1$. Further, any comparisons involving B_2 would be in the

comparison cluster $k = 1$.

With this extension, we let

$$\Gamma_{ij}|\Delta_{ij} = 1, j \in S_k \sim iid\ G_{1k} \tag{3.3}$$

$$\Gamma_{ij}|\Delta_{ij} = 0, j \in S_k \sim iid\ G_{0k}, \tag{3.4}$$

where Φ_{1k} is the parameter vector for G_{1k} and Φ_{0k} is parameter vector for G_{0k}, where $k \in (1, ..., K)$. Although we assume that fields in file B determine the cluster, we could assume that fields in file A determine the cluster without loss of generality. In that case, records in A would be split into sets S_k, and for each pair, the index i would determine the cluster k.

It is possible that the comparison distributions differ only for a subset of the fields. Going back to the address example, we may believe that people between the ages of 18 and 40 move more frequently than the general population. In that case, if we are linking on address, we may consider Frodo Baggins, unknown age, of the Shire, in file A to correspond to 33 year old Frodo Baggins of Buckland in file B if other linking variables are in agreement. We may not expect all fields to follow this pattern, however. If the two Frodos are the same person, the distribution of agreement for fields like gender, birth place, and political affiliation may not be the same as those of a 60 year old woman, and we would want to use the information from other age groups and genders to contribute in our inference on that agreement. Therefore, we split our comparison fields into two groups–those where the inference is shared across linking clusters (e.g., birth place, gender, etc), $f = (1, ..., F_0)$, and those that are different (e.g., address), $f = (F_0 + 1, ..., F)$.

3.3 Model specification

In this section, we outline the BRACS model using the notation and framework from Chapter 3.2.

3.3.1 Likelihood

We model the distributions of the field comparisons by elaborating on (3.3) and (3.4). Let Φ_{1k} be the parameters for G_{1k}, and Φ_{0k} be the parameters for G_{0k}. Let $P_{1k}(\gamma_{ij}|\Phi_{1k}) = P(\Gamma_{ij} = \gamma_{ij}|\Delta_{ij} = 1, \Phi_{1k}, j \in S_k)$, and $P_{0k}(\gamma_{ij}|\Phi_{0k}) = P(\Gamma_{ij} = \gamma_{ij}|\Delta_{ij} = 0, \Phi_{0k}, j \in S_k)$. Let Γ be the set of Γ_{ij} for all $i \in (1, ..., N_A)$ and all $j \in (1, ..., N_B)$. Similarly, let γ be the set of γ_{ij} for all $i \in (1, ..., N_A)$ and all $j \in (1, ..., N_B)$. Because of the iid distributions, and assuming independence of Γ_{1k} and Γ_{0k} given Δ, we can write the likelihood as

$$P(\Gamma = \gamma|\Phi_{0k}, \Phi_{1k}) = \prod_{(i,j)} P_{1k}(\gamma_{ij}|\Phi_{1k})^{\Delta_{ij} * \mathbb{1}(j \in S_k)} P_{0k}(\gamma_{ij}|\Phi_{0k})^{(1-\Delta_{ij}) * \mathbb{1}(j \in S_k)}. \tag{3.5}$$

With conditional independence on the field comparisons given (Δ, S_k), for each field f we have

$$\begin{aligned} P_{1k}(\Gamma^f_{ij} = \gamma^f_{ij}|m_f, j \in S_k) &= \prod_{l=0}^{L_f} (m^k_{fl})^{\mathbb{1}(\gamma^f_{ij}=l) * \mathbb{1}(j \in S_k)} \\ P_{0k}(\Gamma^f_{ij} = \gamma^f_{ij}|u_f, j \in S_k) &= \prod_{l=0}^{L_f} (u^k_{fl})^{\mathbb{1}(\gamma^f_{ij}=l) * \mathbb{1}(j \in S_k)}, \end{aligned} \tag{3.6}$$

where m and u are the parameters in Φ. Thus, m^k_{fl} is the probability that a true link in cluster k has the level of agreement l for field f. Similarly, u^k_{fl} is the probability that a non-link in cluster k has the level of agreement l for field f. As outlined in Chapter 3.2, we assume that for fields $(1, ..., F_0)$, $m^k_f = m_f$ for all records, regardless of S_k. We have

$$\begin{aligned} P_{1k}(\Gamma_{ij} = \gamma_{ij}|\Phi_{1k}, j \in S_k) &= \prod_{f=1}^{F_0} \prod_{l=1}^{L_f} (m_{fl})^{\mathbb{1}(\gamma^f_{ij}=l)} \prod_{f=F_0+1}^{F} \prod_{l=1}^{L_f} (m^k_{fl})^{\mathbb{1}(\gamma^f_{ij}=l)\mathbb{1}(j \in S_k)} \\ P_{0k}(\Gamma_{ij} = \gamma_{ij}|\Phi_{0k}, j \in S_k) &= \prod_{f=1}^{F_0} \prod_{l=1}^{L_f} (u_{fl})^{\mathbb{1}(\gamma^f_{ij}=l)} \prod_{f=F_0+1}^{F} \prod_{l=1}^{L_f} (u^k_{fl})^{\mathbb{1}(\gamma^f_{ij}=l)\mathbb{1}(j \in S_k)}. \end{aligned} \tag{3.7}$$

Thus, we can replace P_{1k} and P_{0k} in (3.5) with those in (3.7), expressing the likelihood as

$$P(\Gamma = \gamma|\Phi_{0k}, \Phi_{1k}) = \prod_{(i,j)} \Bigg[\left(\prod_{f=1}^{F_0} \prod_{l=1}^{L_f} (m_{fl})^{\mathbb{1}(\gamma_{ij}^f = l)} \prod_{f=F_0+1}^{F} \prod_{l=1}^{L_f} (m_{fl}^k)^{\mathbb{1}(\gamma_{ij}^f = l)\mathbb{1}(j \in S_k)} \right)^{\Delta_{ij}} \times \left(\prod_{f=1}^{F_0} \prod_{l=1}^{L_f} (u_{fl})^{\mathbb{1}(\gamma_{ij}^f = l)} \prod_{f=F_0+1}^{F} \prod_{l=1}^{L_f} (u_{fl}^k)^{\mathbb{1}(\gamma_{ij}^f = l)\mathbb{1}(j \in S_k)} \right)^{1-\Delta_{ij}} \Bigg]. \tag{3.8}$$

We assume that any missing data in the linking fields are missing completely at random, and we only include observed fields in the likelihood.

3.3.2 Prior Distributions

We assume a flat prior distribution on the space of Z. In other words, any option for Z_i in our comparison space is equally likely a priori. We enforce that each record in A can link to only one record in B, but for simplicity, we allow multiple records in A to link to the same record in B. We model according to this structure for two reasons. First, this structure makes computation much simpler, which is especially helpful with large data sets. Second, in the NC voting example motivating the algorithm, we care primarily about A, in which there possibly may be duplicates. Ultimately, we find that, in our example, multiple records in A link to the same record in B in only one case, which has minimal effect on the results.

For the parameters Φ_{1k} and Φ_{1k}, we use

$$\begin{aligned} m_f^k &\sim \text{Dirichlet}(\alpha_{0f}^k) \\ u_f^k &\sim \text{Dirichlet}(\beta_{0f}^k), \end{aligned} \tag{3.9}$$

where α_{0f}^k is a vector of length l_f and β_{0f}^k is a vector of length l_f. We discuss how to set α_{0f}^k and β_{0f}^k in Chapter 3.4.

3.3.3 Gibbs Sampler

In this section, we outline the steps for the Gibbs sampler we use to draw from the posterior distributions of Z, Φ_{0k}, and Φ_{1k}.

3.3.4 Sampling Φ_{0k} and Φ_{1k}

For each iteration t, and for $i = 1, ..., N_b$, where $f = (1, ..., F_0)$, sample

$$\begin{aligned} m_f^{(t)}|\gamma, z^{(t-1)} &\sim \text{Dirichlet } (\alpha_{0fl} + \sum_{ij} \mathbb{1}(\Delta_{ij}^{(t-1)} = 1)\mathbb{1}(\gamma_{ij}^f = l)) \\ u_f^{(t)}|\gamma, z^{(t-1)} &\sim \text{Dirichlet } (\beta_{0fl} + \sum_{ij} \mathbb{1}(\Delta_{ij}^{(t-1)} = 0)\mathbb{1}(\gamma_{ij}^f = l)). \end{aligned} \tag{3.10}$$

Similarly, for each iteration t, and for $i = 1, ..., N_b$, where $f = (F_0 + 1, ..., F)$, sample

$$\begin{aligned} m_f^{k(t)}|\gamma, S_k, z^{(t-1)} &\sim \text{Dirichlet } (\alpha_{0fl}^k + \sum_{ij} \mathbb{1}(\Delta_{ij}^{(t-1)} = 1)\mathbb{1}(\gamma_{ij}^f = l) * \mathbb{1}(j \in S_k)) \\ u_f^{k(t)}|\gamma, S_k, z^{(t-1)} &\sim \text{Dirichlet } (\beta_{0fl}^k + \sum_{ij} \mathbb{1}(\Delta_{ij}^{(t-1)} = 0)\mathbb{1}(\gamma_{ij}^f = l) * \mathbb{1}(j \in S_k)). \end{aligned} \tag{3.11}$$

3.3.5 Sampling Z

To begin, we note that

$$P(Z_i = j|\Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, \gamma, j \in S_k) = c_j \times P(\gamma_{ij}|Z, \Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) \tag{3.12}$$

where c_j is a normalizing constant. We see that, in cases where Z_i has a match, i.e., $Z_i = j$ for some $j \in (1, ..., N_B)$,

$$P(\gamma_{ij}|Z, \Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) = \prod_{f=1}^{F} \prod_{l=1}^{L_f} m_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij}^f = l)}} \times \prod_{j' \neq j} \prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij'}^f = l)}} \tag{3.13}$$

where j' indicates the other indices in $j = (1, ..., N_B)$. This is equivalent to

$$P(\gamma_{ij}|Z, \Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) = \prod_{f=1}^{F}\prod_{l=1}^{L_f} m_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij}^f=l)}} \times \prod_{j'\neq j}\prod_{f=1}^{F}\prod_{l=1}^{L_f} u_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij'}^f=l)}} \times \frac{\prod_{f=1}^{F}\prod_{l=1}^{L_f} u_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij}^f=l)}}}{\prod_{f=1}^{F}\prod_{l=1}^{L_f} u_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij}^f=l)}}}. \tag{3.14}$$

Thus, for $j \leqslant N_b$,

$$P(Z_i = j|\Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) = \prod_{f=1}^{F}\prod_{l=1}^{L_f} \left(\frac{m_{fl}^{k(t)}}{u_{fl}^{k(t)}}\right)^{\mathbb{1}(\gamma_{ij}^f=l)} \times c_j \times c_{j'} \tag{3.15}$$

where $c_{j'} = \prod_{j=1}^{N_B}\prod_{f=1}^{F}\prod_{l=1}^{L_f} u_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij}^f=l)}}$.

For $j = N_b + 1$, we calculate

$$P(\gamma_{ij}|Z, \Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) = \prod_{j=1}^{N_B}\prod_{f=1}^{F}\prod_{l=1}^{L_f} u_{fl}^{k(t)^{\mathbb{1}(\gamma_{ij}^f=l)}} \tag{3.16}$$

which is equivalent to $c_{j'}$.

Thus,

$$\begin{aligned} &P(Z_i = j|\Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) = c_j \times c_{j'} \times \prod_{f=1}^{F}\prod_{l=1}^{L_f} \left(\frac{m_{fl}^{k(t)}}{u_{fl}^{k(t)}}\right)^{\mathbb{1}(\gamma_{ij}^f=l)} \text{ for } j \leqslant N_B, \\ &P(Z_i = j|\Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) = c_j \times c_{j'} \text{ for } j = N_B + 1. \end{aligned} \tag{3.17}$$

So for each iteration t, we sample:

$$\begin{aligned} &P(Z_i = j|\Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) \propto \prod_{f=1}^{F}\prod_{l=1}^{L_f} \left(\frac{m_{fl}^{k(t)}}{u_{fl}^{k(t)}}\right)^{\mathbb{1}(\gamma_{ij}^f=l)} \text{ for } j \leqslant N_B, \\ &P(Z_i = j|\Phi_{1k}^{(t)}, \Phi_{0k}^{(t)}, j \in S_k) \propto 1 \text{ for } j = N_B + 1. \end{aligned} \tag{3.18}$$

We can therefore sample from the posterior distribution of Z_i by normalizing multinomial probabilities equal to $\prod_{f=1}^{F}\prod_{l=1}^{L_f}(\frac{m_{fl}^{k(t)}}{u_{fl}^{k(t)}})^{\mathbb{1}(\gamma_{ij}^f=l)}$ for each $j \leqslant N_B$, and equal to 1 for $j = N_B + 1$.

3.4 Simulation: Evaluation with low baseline match rates

In the simulations in this section, we imitate possible structure and behavior of the files we link in the NC voting application. We use entirely simulated data to evaluate how the proposed method compares to a non-clustered model in situations where (i) there is an underlying clustered structure and (ii) there is not an underlying clustered structure.

In order to mimic a collection of two datasets, we generate identification numbers (IDs) $A_{i0} \in (1, ..., 200)$ for A, representing the list of provisional voters, and $B_{j0} \in (101, ..., 1000)$ for the larger B, representing the voter file. Thus, records with IDs between 101 and 200 are links across the two files. We then generate comparison data for 7 of 8 comparison fields representing comparisons of city, gender, political party affiliation, first name, middle name, and last name. The name comparisons are based on the name comparison distributions in the *RecordLinkage* R package data (Borg and Sariyar, 2015), which we use in later simulations.

In the first two simulations, our comparison distributions Γ_f for the first seven fields are drawn as shown in Table 3.1. In the table, "Field" refers to the comparison field, Γ_f is the comparison vector for that field, and the values in columns $\Delta_{ij} = 1$ and $\Delta_{ij} = 0$ are the distributions of the Γ_f where the pair (A_i, B_j) is and is not a link, respectively.

We additionally simulate an eighth comparison field representing address which varies by k in the simulation in Chapter 3.4.1 and does not vary in the simulation in Chapter 3.4.2.

Table 3.1: Comparison distributions for simulated data in Chapter 3.4.

Field	Γ_f	$\Delta_{ij} = 1$	$\Delta_{ij} = 0$
City	Γ_1	Bern(.9)	Bern(.2)
Gender	Γ_2	Mult(.05, .05, .9)	Mult(.5, .3, .2)
Race	Γ_3	Mult(.05, .05, .9)	Mult(.7, .2, .1)
Party	Γ_4	Mult(.05, .15, .8)	Mult(.9, .05, .05)
First name	Γ_5	Mult(.005, .295, .05, .65)	Mult(.97, .01, .01, .01)
Middle name	Γ_6	Mult(.01, .24, .05, .70)	Mult(.97, .01, .01, .01)
Last name	Γ_7	Mult(.05, .05, .20, .70)	Mult(.925, .025, .025, .025)

We use the same hyper-parameters in the simulations and in the application in Chapter 4. The values of α_0^k and β_0^k are listed in Table 3.2. One can interpret these priors as weights we give to our beliefs about the distributions. For example, our prior for gender agreeing is Dirichlet$(3, 2, 1)$ under non-matches. We are essentially giving our prior the weight of six observed units that are non-matches. Three disagree completely on gender, two neither agree nor disagree, and one agrees completely. We give more weight to the string field priors, particularly for matches. We feel confident that true links are very likely to have the same name, with room for typographical errors, across the two files. While we do not generally find large variations in the results due to the particular values of the priors, we do see that more informative priors can give us better results. This is a case where we can use our prior knowledge and expectations to inform the parameters, as in the gender example, and the results suggest that this subjective Bayesian approach improves the accuracy of the linkage. Informative priors help guide the model to a reasonable posterior mode without overpowering the signal in the data, e.g., not all matches or non-matches, as we are very confident in the presence of at least some underlying matches and some underlying non-matches in the data.

We perform record linkage using the proposed model and the Gibbs sampler steps and priors outlined in Chapter 3.3, with variation in the eighth field. For

Table 3.2: Values of α_0^k and β_0^k as hyper-parameters for m and u priors

Field	α_0^k	β_0^k
City	(2, 1)	(1, 20)
Gender	(3, 2, 1)	(1, 5, 10)
Race	(3, 2, 1)	(1, 5, 10)
Party	(3, 2, 1)	(1, 5, 10)
First name	(10, 5, 2, 1)	(1, 10, 50, 100)
Last name	(10, 5, 2, 1)	(1, 10, 50, 100)
Middle name	(10, 5, 2, 1)	(1, 10, 50, 100)
Address	(10, 5, 2, 1)	(1, 10, 50, 100)

comparison, we use a model with the same steps and priors but only one m_8 and u_8 and therefore not accounting for a clustered structure. For both approaches, we generate a field representing Levenshtein name similarities for all pairs and consider those with similarities below 0.5 to be non-links. The Levenshtein name similarity cutoff of 0.5 limits the comparison space for computational purposes. It is intended to allow for typographical mistakes, but may exclude some nicknames or name changes. We perform 5,000 MCMC iterations with a burn-in 1,000 burn-in samples.

We compare posterior match identification rates and non-match identification rates for the two approaches. We define match identification rates as the proportion of the 100 true matches in A that are correctly identified in each post burn-in draw of Z, or $\frac{1}{100}\sum_{i=101}^{200}\mathbb{1}(Z_i = A_{i0})$ and non-match identification rates as the proportion of the 100 non-matches in file A that are correctly classified as non-matches, or $\frac{1}{100}\sum_{i=1}^{100}\mathbb{1}\big(Z_i = (N_B + 1)\big)$.

3.4.1 Clustered structure: one field

In this simulation, we draw linking variables 1-7 from the distributions outlined in the beginning of Chapter 3.4. Additionally, we link on a field 8 with a structure that depends on attributes in file B, such that we know cluster k based on the index j.

We generate field comparisons for this eighth linking variable, meant to mimic

Table 3.3: Address comparison distributions for simulated data in Chapter 3.4.1.

S_k	$\Delta_{ij} = 1$	$\Delta_{ij} = 0$
S_1	Mult(.75, .1, .1, .15)	Mult(.9, 05, .025, .025)
S_2	Mult(.6, .1, .15, .15)	Mult(.8, 15, .025, .025)
S_3	Mult(.5, .1, .2, .2)	Mult(.85, 05, .05, .05)
S_4	Mult(.2, .1, .3, .4)	Mult(.7, .1, .1, .1)

address in our applied analysis, as shown in Table 3.3. In the table, S_k indicates the cluster for the pair (A_i, B_j). The values in columns $\Delta_{ij} = 1$ and $\Delta_{ij} = 0$ are the distributions of the $\Gamma_{address}$ where the pair (A_i, B_j) is and is not a link, respectively. "Mult" refers to the multinomial distribution.

The distribution for $\Gamma_{address}$ depends on both Δ and S_k such that the comparisons have stark differences among links and non-links for higher k. We choose these multinomial probabilities because we expect to see a similar distribution among address comparisons with age-defined clusters. Older people may move less often (Mateyka, 2015), and so if we place records for the oldest people in S_4 based on an age field in B, address will carry more weight as a comparison than it will for younger people.

In this simulation, the posterior match identification rate is 0.98 for BRACS and where we do not account for the clustered structure. The non-match identification rate varies more, both within and across the two approaches. Figure 3.1 shows the posterior distributions of the non-match identification rates under the two models. While BRACS identifies, on average, approximately 85% of the non-matches, that number drops to less than 60% when we do not account for the clustered structure.

3.4.2 Non-clustered structure

For this simulation, we generate our eighth field comparison as follows:

$$\Gamma_8|\Delta_{ij} = 1 \sim \text{Mult}(.5, .1, .2, .2),\ \Gamma_8|\Delta_{ij} = 0 \sim \text{Mult}(.85, .05, .05, .05).$$

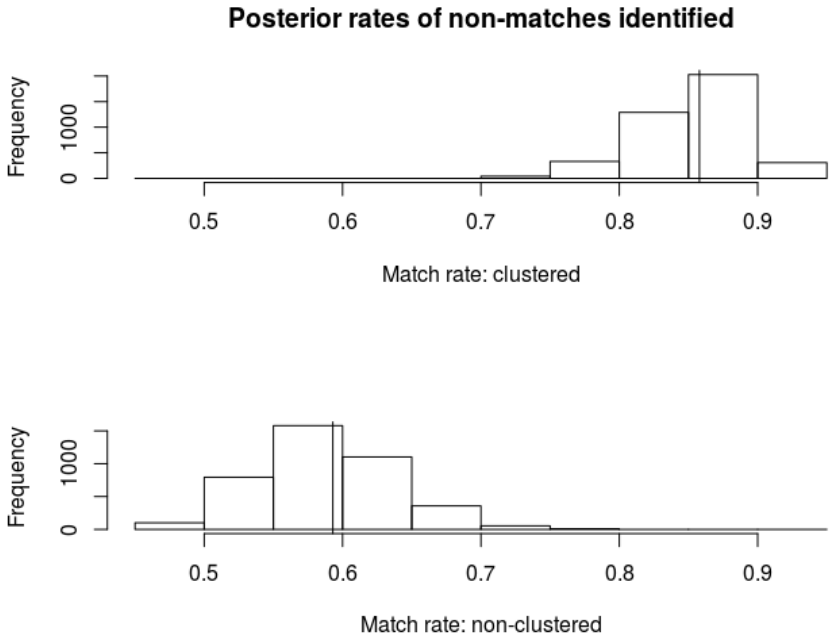

FIGURE 3.1: Posterior rate of non-match identification for the simulation in Chapter 3.4.1, accounting for and not accounting for the underlying clustered structure.

The distribution, therefore, does not necessitate the use of BRACS.

Similar to Chapter 3.4.1, we see consistently high match identification rates (0.99 for both approaches). We also again see some variation in non-match identification rates. Interestingly, BRACS slightly out-performs the existing model even though the data are not generated according to the clustered structure, as we show in Figure 3.2. We believe that this is because the strong priors push the model away from identifying false-positives, and the increase in the number of parameters increases the strength of the model assumptions implied by the priors.

3.5 Simulation studies: RecordLinkage R package data

In this section, we again use simulated data to evaluate how BRACS compares to a non-clustered model. We simulate situations in which (i) two linkage fields have underlying clustered structures, and (ii) there is no underlying clustered structure.

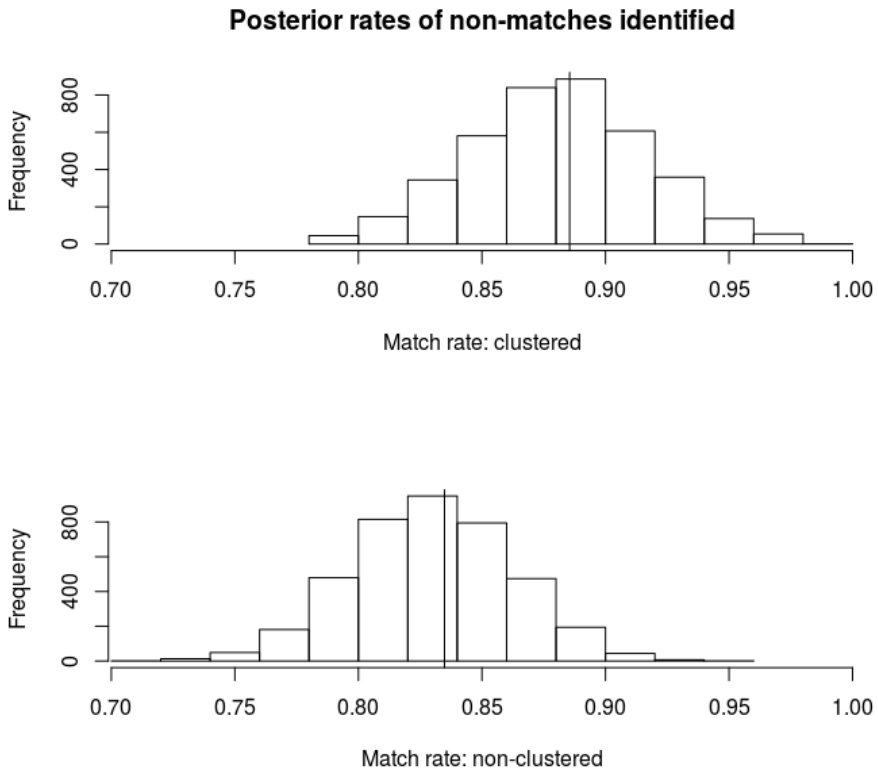

FIGURE 3.2: Posterior rate of non-match identification for the non-clustered simulation in Chapter 3.4.2, using BRACS vs. a non-clustered model.

For these two simulations, rather than using entirely simulated data, we first use the RL10000 dataset from the recordLinkage package in R (Borg and Sariyar, 2015). The package has two datasets to test record linkage methods, one of which is called RL10000. It has 8,000 unique records, plus 1,000 records which appear twice but with errors introduced. Name and birth date (birth year, birth month, and birth day) are all included in the data set. To make the data appropriate for linkage rather than de-duplication, we use the identifiers on the file to split the 1,000 duplicated records into two sets, A and B. We then append 1,000 of the non-duplicated records to A and the remaining 8,000 to B. Next, in order to block on birth year, we modify the birth year for the duplicated records in B so that they match those in A. Finally, we add two additional linking fields, representing a hypothetical address comparison

Table 3.4: Address comparison distributions for simulated data in Chapter 3.5.1.

S_k	$\Delta_{ij} = 1$	$\Delta_{ij} = 0$
S_1	Mult(.75, .05, .05, .15)	Mult(.9, 05, .025, .025)
S_2	Mult(.6, .1, .15, .15)	Mult(.8, 15, .025, .025)
S_3	Mult(.5, .1, .2, .2)	Mult(.85, 05, .05, .05)
S_4	Mult(.2, .1, .3, .4)	Mult(.7, .1, .1, .1)

based on age, generating field comparisons which allow us to test how the method performs under the different scenarios.

We perform record linkage using the proposed model and the Gibbs sampler steps and priors outlined in Chapter 3.3 for names and our simulated address comparison field, plus $Gamma(2, 1)$ and $Gamma(1, 20)$ priors for the u and m parameters for birth month and birth day. To compare our method to a non-clustered approach, we use the same steps and priors but with only one $m_{address}$ and $u_{address}$ and therefore no underlying clustered structure.

For both approaches in the following three simulations, we consider all pairs with Levenshtein name similarities below 0.5 to be non-links. We believe this cutoff will pick up most nicknames and typos, but not completely different names (e.g., JOAN and JODY has a Levenshtein similarity of 0.5, but JACOB and JODY has a similarity of 0.2). We perform 5,000 iterations with a burn-in 1,000 burn-in samples.

3.5.1 Linkage with variation in two fields

In this simulation, we generate one address comparison as follows according to Table 3.4. In the table, S_k indicates the cluster for the pair (A_i, B_j). The values in columns $\Delta_{ij} = 1$ and $\Delta_{ij} = 0$ are the distributions of the $\Gamma_{address}$ where the pair (A_i, B_j) is and is not a link, respectively. "Mult" refers to the multinomial distribution.

Similar to Chapter 3.4.1, the distribution for $\Gamma_{address}$ depends on both Δ and S_k, with more differences between links and non-links for higher k.

Next, we generate an additional field with a different age-dependency (it could

Table 3.5: Email comparison distributions for simulated data in Chapter 3.5.1.

S_k	$\Delta_{ij} = 1$	$\Delta_{ij} = 0$
S_1	Mult(.2, .1, .3, .4)	Mult(.7, .1, .1, .1)
S_2	Mult(.5, .1, .2, .2	Mult(.85, 05, .05, .05)
S_3	Mult(.6, .1, .15, .15)	Mult(.8, 15, .025, .025)
S_4	Mult(.75, .05, .05, .15)	Mult(.9, 05, .025, .025)

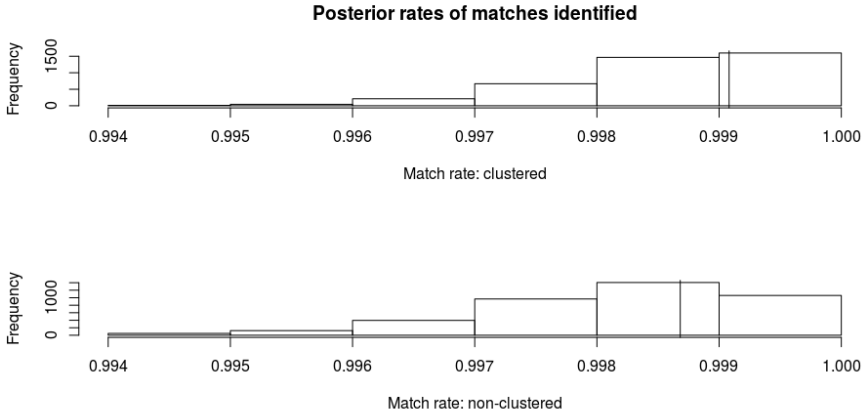

FIGURE 3.3: Posterior rate of match identification for the simulation in Chapter 3.5.1, using BRACS vs. a non-clustered model.

be, for example, an email address comparison) according to Table 3.5. In the table, S_k indicates the cluster for the pair (A_i, B_j). The values in columns $\Delta_{ij} = 1$ and $\Delta_{ij} = 0$ are the distributions of the Γ_{email} where the pair (A_i, B_j) is and is not a link, respectively. "Mult" refers to the multinomial distribution.

Figures 3.3 and 3.4 show that both approaches nearly always correctly identify all of the links, but using BRACS yields gains in non-match identification rates. Because the baseline match and non-match identification rates are so high, there is not as much room for improvement in this example as in Chapter 3.4.1. Therefore the marginal gains are smaller.

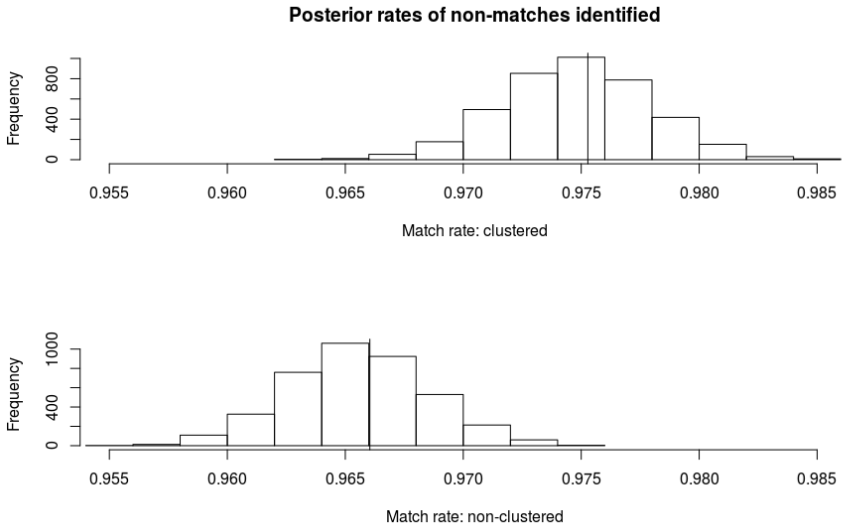

FIGURE 3.4: Posterior rate of non-match identification in Chapter 3.5.1, using BRACS vs. a non-clustered model.

3.5.2 Linkage with non-clustered structure

In this simulation, linking variables 1-7 have relationships drawn from the distributions outlined in the beginning of Chapter 3.5. We also link on a hypothetical address field, for which comparisons are drawn as follows.

$$\Gamma_{address}|\Delta_{ij} = 1 \sim \text{Multinom}(.5, .1, .2, .2),$$
$$\Gamma_{address}|\Delta_{ij} = 0 \sim \text{Multinom}(.85, .05, .05, .05).$$

This simulation examines the performance of the algorithm when there are no between-cluster differences in the distribution of $\Gamma_{address}$.

Figures 3.5 and 3.6 show posterior distributions of the match identification rates and non-match identification rates. The results of the two approaches are essentially identical, indicating that using the model in the absence of the underlying clustered structure is not necessarily detrimental to the linkage identification. However, there are more parameters to estimate, so the increase in variance due to parameter estimation may lead to decreased quality in performance in other contexts.

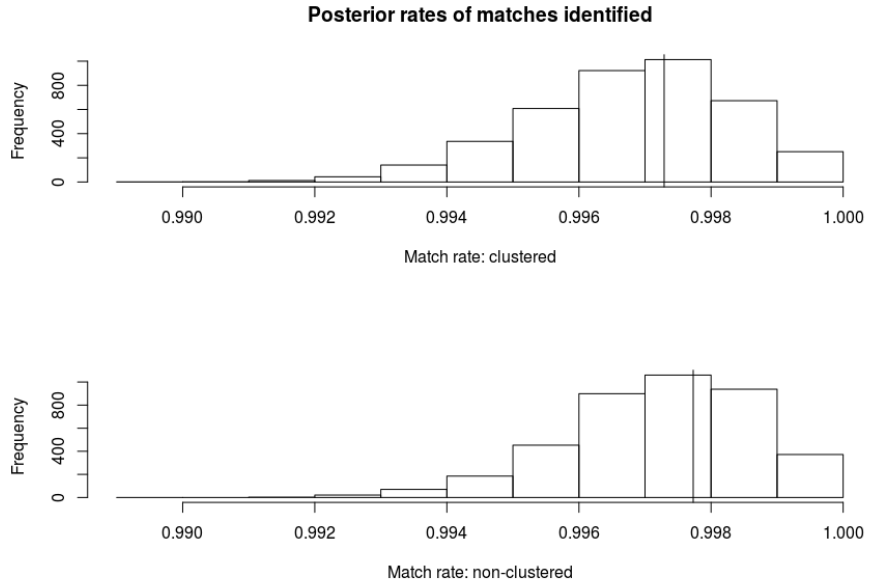

FIGURE 3.5: Posterior rate of match identification for the simulation in Chapter 3.5.2, accounting for and not accounting for a clustered structure in one field agreement.

3.6 Conclusions

The results from the simulation studies suggest that clustered linkage can improve linkage quality when Γ differs by group. In both simulation scenarios, we saw the most improvement in the identification of non-matches. When the distribution of an agreement field, such as address, varies in a systematic way, the m and u parameters will not reflect the agreement probabilities appropriately. By allowing the m and u parameters to change by group, we allow for more precise estimation of the likelihood of each pair being a match, and thus higher classification rates.

When Γ does not differ by group, adding dependency on S_k may add noise in the estimates due to smaller sample sizes available to estimate the parameters m_k and u_k. It may also make linkage more sensitive to a strong prior. These strengthened assumptions can help accuracy, as there are certain contexts (including voter file linkage) in which there has been substantial work previously done, and prior

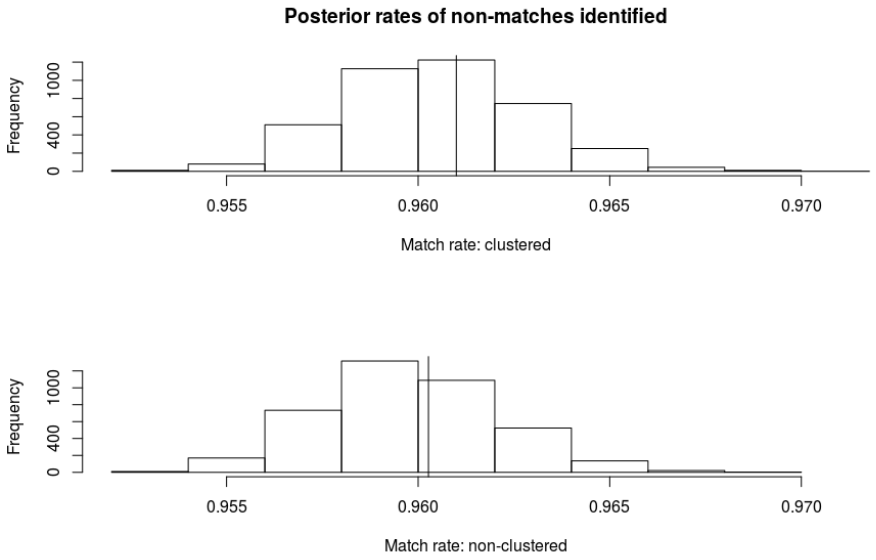

FIGURE 3.6: Posterior rate of non-match identification for the simulation in Chapter 3.5.2, accounting for and not accounting for a clustered structure in one field agreement.

information may be both useful and easily obtained.

There can be increased variance in using BRACS, but there can be substantial gains in accuracy, particularly when there is large variation in comparison distributions conditional on k, or when matches or non-matches are difficult to identify. In addition to the gains in accuracy, the Bayesian nature of the model also allows us to account for linkage uncertainty, which can then be used in further analytic decisions. These may include further modeling of voter behavior, or simply voter identification for follow-up. The posterior distribution provided by this approach allows for this post-linkage flexibility.

There are challenges inherent in the approach as well. Adding additional parameters increases the computational complexity, which is already a challenge for Bayesian record linkage approaches. We mitigate this problem by imposing additional restrictions to our comparison space based on name, but depending on context, the additional complexity due to the additional parameters may not be worth

the computational cost. Additionally, there is a challenge in deciding which fields to condition on. In some cases, there may be built-in categorizations (like age group on the voter file) and contextual knowledge to guide the decision. However, the complexity increases non-linearly with the number of conditional fields. How to account for higher numbers of conditional fields is a topic for future research.

4

BRACS: Application to North Carolina Voting Data

4.1 Introduction

We developed a method for Bayesian Record linkage And Clustered Sub-models (BRACS), outlined in Chapter 3, intended for linking the list of 2016 North Carolina provisional voters and the list of registered voters in North Carolina as of November 8, 2016 (the voter file). This chapter outlines the two data sets and the results of applying BRACS to link them together. We structure the chapter as follows. In Chapter 4.2, we describe provisional voting and outline the data and structure of the publicly accessible file listing provisional voters. In Chapter 4.3, we describe the publicly available North Carolina voter file. In Chapter 4.4, we describe limitations in the data that complicate linking the two files together. In Chapter 4.5, we use BRACS to link the two files together. In Chapter 4.6, we present results of this analysis without using information provided after the provisional ballot canvass, mimicking the scenario that political campaigns face in contested elections. We provide posterior intervals for how these results compare to a simple join with IDs that

are later provided by the state. We conclude this chapter section with a discussion on how BRACS can lead to improved match rates above a non-Bayesian approach, and how the posterior distribution may be used to prioritize voter contact in a campaign that works with voters to provide additional information required to make their votes count.

In Chapter 4.7, we use BRACS to estimate the number of provisional voters that had been removed from the voter file prior to the election. Because of the limitations of the state-provided identifiers, we use those identifiers as linking variables in the analysis, rather than as ground-truth. However, we show posterior comparison intervals, similar to Chapter 4.6, as a reference. We also provide posterior intervals of the estimated number of removed voters.

We conclude with a discussion of the results and of the linkage method it applies to the data application.

4.2 Provisional Voting

4.2.1 Background

When attempting to vote, some voters are directed to cast a provisional ballot. These ballots are required under various circumstances, including, but not limited to, incomplete voter registration information, attempting to vote in the wrong precinct, an unreported move after registering to vote, or being previously removed from the voter file (see Table 4.1 for tallies of ballot approval status and reasons in this data set). The voter writes down some personal information and the ballot is sealed. After election day, the county Boards of Elections determine which of the ballots should be counted before opening the ballots and tallying the votes; this process is called the canvass. Information on who has cast a provisional ballot is public, and individuals, organizations, or political campaigns may attempt to match these voters to the voter file to determine information pertaining to the likelihood that

the vote is counted and whether additional action is required on behalf of the voter. Provisional ballots play an important role in contested races, a notable example of which is Al Franken's initial Senatorial race (CNN Political Ticker, 2009). North Carolina's 2016 gubernatorial race (Jarvis, 2016) also heavily featured provisional ballots, as the race between incumbent Pat McCrory and challenger Roy Cooper was not called until after the absentee and provisional ballots were counted.

4.2.2 Data Structure

There are 60,643 records in North Carolina's 2016 provisional ballot file, comprising each voter's county of residence, voting date, voting precinct, name, address, city, state, zip code, reason for casting the provisional ballot, eventual status of the vote (e.g., Approved, Not Counted), whether the vote was curbside, reason for the vote not being counted (if applicable), phone number, party, sex, ethnicity, race, and a county-specific voter registration number. Some of the fields, including voter registration number, vote status, and reason for that status, are not available during the approval process but are appended afterward. Table 4.1 shows the approval status and, where relevant, non-approval reasons for North Carolina's 2016 provisional votes.

4.3 North Carolina Voter File

4.3.1 Background

In addition to provisional voter information, the North Carolina Board of Elections makes a considerable amount of voter registration information publicly available, free of charge, for download. The provisional file includes a field indicating whether the vote was counted, and if not, the reason. There is significant county-level variability as well as room for subjectivity in this field, prompting the question of whether it accurately reflects a standardized reason for not counting certain votes. One of the reasons for a vote not to count is that a voter had been removed from the rolls.

Table 4.1: Approval status and non-approval reasons for votes listed on the NC provisional file.

Approval	Not counted reason	Ballots
Approved		21717
Not Counted	Not Registered	26890
Not Counted	Removed	2843
Not Counted	Registration After Deadline	803
Not Counted	ID Not Provided	769
Not Counted	Moved Out of County More Than 30 Days	763
Not Counted	Previously Denied	532
Not Counted	Ineligible to Vote	310
Not Counted	Other	231
Not Counted	Voter Already Voted	182
Not Counted	Not Eligible to Vote In Current Election	120
Not Counted	Ballot Missing from Envelope	113
Not Counted	Provisional Application Incomplete/Illegible	99
Not Counted	Transaction Cancelled	50
Not Counted	Voting Out of Precinct	44
Not Counted	Non-Matching Signature	7
Partial	Voting Out of Precinct	4600
Partial	Other	500
Partial	Not Registered	27
Partial	Removed	21
Partial	Not Eligible to Vote in Current Election	8
Partial	Wrong Party Ballot	6
Partial	Moved Out of County More Than 30 Days	3
Partial	Provisional Application Incomplete/Illegible	3
Partial	Voter Already Voted	2

Removal can happen for a number of reasons, ranging from a felony conviction, to death, to not voting in two federal election cycles and failing to respond to an address verification. This last reason was the subject of a lawsuit in North Carolina in 2016, resulting in a ruling that NC had prematurely purged approximately 6700 voters from the rolls (Williams, 2016). Knowing how many removed voters attempt to vote on Election Day could help judge the impact of such purges, but it necessitates linking the provisional file to the voter file.

4.3.2 Data Structure

North Carolina releases voter file snapshots of the file at a specific point in time. The snapshot for November 8, 2016 has 13,254,628 records. Multiple records may correspond to the same person (across different addresses or name changes), but there is a person-level identifier in the file, called the NCID, to distinguish these duplicates. Each voter registration record is assigned the county-specific voter registration number.

This identifier could be used to link the provisional file to the voter file. However, we believe that this identifier may be inadequate for fully linking the two files together, as we discuss in Chapter 4.4.

4.4 Limitations of the voter registration number as a unique identifier

The voter registration number does not appear until after the provisional ballot canvass, so if a campaign or political group wished to perform record linkage, they would be unable to do so until after the votes had been counted. After it is appended, there are inconsistencies in the use of the IDs, suggesting that they are unreliable for perfect linkage across the two files. For example, out of the 60,643 records in provisional file, 8,875 have a voter registration number listed that is not found in the voter's county on the election date's voter file snapshot. Additionally, there are 101 voters whose votes were not counted with a reason of "Not Registered," but according to their voter registration number listed, were in fact registered and eligible at the time.

These inconsistencies vary significantly by county, as individual county Boards of Elections have some autonomy in record-keeping (North Carolina State Board of Elections, 2017). Table 4.2 breaks down some inconsistencies in the ID-based joins of the provisional file to the voter file for a sample of the counties. In the table,

Table 4.2: Percent inconsistencies in ID-based join by county for a sample of NC counties.

County Name	Total provisional votes	VR number listed not found	Approved with no match
Alamance	1105	8%	9%
Carteret	394	62%	38%
Durham	1926	7%	6%
Forsyth	1881	11%	10%
Franklin	311	76%	17%
Hertford	281	1%	1%
Iredell	726	23%	22%
Mecklenburg	3778	7%	6%
Orange	428	17%	17%
Robeson	2067	3%	3%
Swain	36	94%	47%
Transylvania	216	2%	2%

"Total provisional votes" refers to the number of voters on the provisional file within each county. "VR number listed not found" refers to cases where there is a voter registration number listed on the provisional file, but that number is not found on the voter file in that county. "Approved with no match" refers to "VR number listed not found" cases where the vote was ultimately approved. All counties are listed in Table 7.10 in the chapter appendix.

4.5 Using BRACS to merge the NC voter file and the list of 2016 NC provisional voters

Though we have described limitations of the voter registration number, we performed exploratory analysis on records merged with the ID in order to identify patterns that could complicate the linkage. True matches may be missed in this analysis, but we believe that positive matches according to the voter registration number are highly likely to correspond to the same voter. When looking through these records linked on voter registration number, we found that there were different levels of address agreement based on age groups (see Figure 4.1 for address Levenshtein

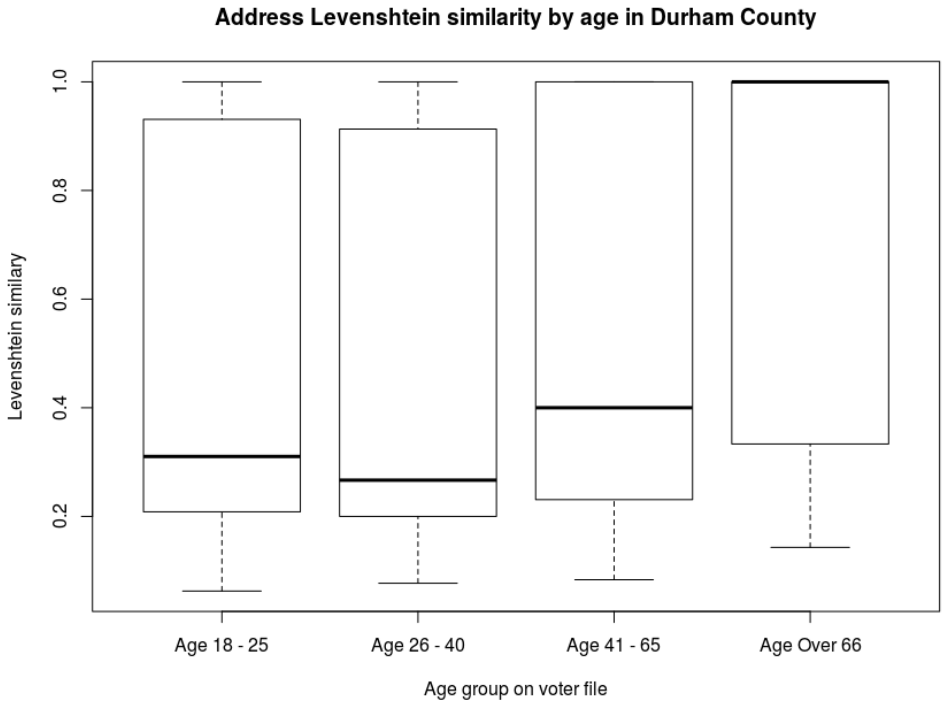

FIGURE 4.1: Durham County address Levenshtein similarity by age: voter registration number matches only

similarity in Durham County among voters in our comparison space, and Figure 7.4 in the appendix for address Jaro-Winkler scores in Durham County among the same voters). Older voters are less likely to have a different address listed between the two files, whereas younger voters (especially those between the ages of 26 and 40) seem more likely to have moved between the time of registration and casting their provisional ballot. These differences indicate that a simple model that treats all address comparisons as coming from the same underlying distribution would be inaccurate. Therefore, we split the address comparisons into the four age categories listed in the NC voter file: Age 18-25, 26-40, 41-65, and Over 66. We initially planned to follow a similar procedure for last name comparisons based on gender, but doing so prevented us from limiting the comparison space based on Levenshtein distance of last names, and the higher computation costs were not feasible on a large scale.

4.5.1 Blocking procedure

Due to the large scale of the data, and particularly the size of the North Carolina voter file, this application necessitates blocking, or only considering record pairs with exact matches on certain fields as potential links, for viable computation. Fewer fields are required on the provisional file than on the voter file, and the provisional file is less standardized, meaning there are no truly reliable blocking fields. However, if the county that a person votes in is not the same as the county on his or her voter registration record, the vote will not count. Therefore, if we do not consider registered voters in Durham county as possible links for a provisional voter in Orange county, any unidentified true links will not have votes counted and so are of minimal importance for a political campaign application. Given these considerations, we use county as a blocking variable.

Even blocking on county, there is a scale issue. Wake county, for example, has 1,269,231 voter file records and 6,793 provisional records. Considering each record pair would mean 8,621,886,183 comparison pairs for Wake alone. Therefore, we decided to use a cutoff for name similarity in addition to the county blocks. We consider only pairs with a normalized Levenshtein similarity of first names and last names above 0.7, as well as a normalized Levenshtein similarity of middle names above 0.25 as potential matches, although we include pairs with an average first/last Levenshtein similarity above 0.7 in calculations of the u_f.

These additional criteria greatly reduce the comparison space. We create the county-specific blocks separately, partly for computational purposes, as this allows for parallel computing, but also because the votes are counted and recorded at the county level, so we expect inter-county variability in the linkage parameters.

We code the γ_{ij} as follows. If $\text{city}_i \neq \text{city}_j$, $\gamma_{ij}^{city} = 0$. If $\text{city}_i = \text{city}_j$, $\gamma_{ij}^{city} = 1$. If $\text{gender}_i = \text{gender}_j$, $\gamma_{ij}^{gender} = 2$. If $\text{gender}_i = \text{M}$ and $\text{gender}_j = \text{F}$ or $\text{gender}_i =$

F and gender$_j$ = M, $\gamma_{ij}^{gender} = 0$. Otherwise, $\gamma_{ij}^{gender} = 1$. This three-level split allows for cases when someone marks their gender, but marks it as undesignated or other. Similarly, if race$_i$ = race$_j$, $\gamma_{ij}^{race} = 2$. If race$_i$ = B and race$_j$ = W or race$_i$ = W and race$_j$ = B, $\gamma_{ij}^{race} = 0$. Otherwise, $\gamma_{ij}^{race} = 1$. "B" (Black) and "W" (White) are the most common races listed on the provisional file, and occurrences of self-reported race switching between White and Black across the two files appear very rare, whereas switches to "Undesignated" or "Other" are more common. For party affiliation, we follow the three-level split again, with Democrat/Republican switches occupying their own category. Thus, if party$_i$ = party$_j$, $\gamma_{ij}^{party} = 2$. If party$_i$ = R and party$_j$ = D, or party$_i$ = D and party$_j$ = R, we set $\gamma_{ij}^{party} = 0$. Otherwise, $\gamma_{ij}^{party} = 1$. We split first name, middle name, last name, and address using the following Levenshtein similarity cutoffs. Let L_{ij}^f be the similarity between f_i, f_j where f is either first name, middle name, last name, or address. We set $\gamma_{ij}^f = 0$ if $L_{ij}^f < 0.7$, $\gamma_{ij}^f = 1$ if $0.7 \leqslant L_{ij}^f < 0.9$, $\gamma_{ij}^f = 2$ if $0.9 \leqslant L_{ij}^f < 1$, and $\gamma_{ij}^f = 3$ if $L_{ij}^f = 1$, indicating an exact match. Middle names often have only one initial listed, and so if one of the comparison fields has only one character, we look at agreement between only the first characters in the middle name string, in order to give more weight to the first initial match.

We perform the linkage according to the steps outlined in Chapter 3.3. We report the posterior mode for each Z_i after running an MCMC chain of length 5000 and a burn-in of length 1000. We compare the results to a simple join using only the IDs listed in the provisional file.

4.6 Matching provisional voters to the file using linkage without state-provided IDs

In this chapter section, we link the list of provisional voters to the voter file without incorporating voter registration numbers. This scenario mimics the analysis that would be performed between election day and the canvass. A benefit of the Bayesian approach in this scenario is that one could prioritize people to call based on posterior distributions of match rates. For example, one could calculate the posterior probability of each potential match pair and rank potential matches to the provisional file until the cumulative probabilities for each voter was at or above some threshold p_m. Then individuals associated with the local Board of Elections could attempt to contact people on that list until the identity was resolved. If a campaign was making follow-up calls, one could further prioritize based on the likelihood of the match as well as the likelihood of the matched voter being a supporter of the candidate. The integration of a formalized decision rule to rank potential matches is an interesting area for future research.

To summarize the results, we examine agreement rates for groups defined by the state-provided IDs. Specifically, figures 4.2 and 4.3 show the posterior classification agreement of pairs categorized by the state provided identifiers as model matches and model non-matches, respectively. We define the classification agreement of matches at a given iteration t as $\frac{1}{n_m} \sum_{(i,j) \in C_m} \Delta_{i,j}^t$ where n_m is the number of matches according to the voter registration number, and C_m is the comparison space, limited to record pairs classified as matches by the voter registration number.

Similarly, we define the classification agreement of non-matches at a given iteration t as $1 - \frac{1}{n_p - n_m} \sum_{(i,j) \in C'_m} \Delta_{i,j}^t$ where n_p is the number of provisional voters to link, and C'_m is the comparison space limited to record pairs *not* classified as matches by the voter registration number.

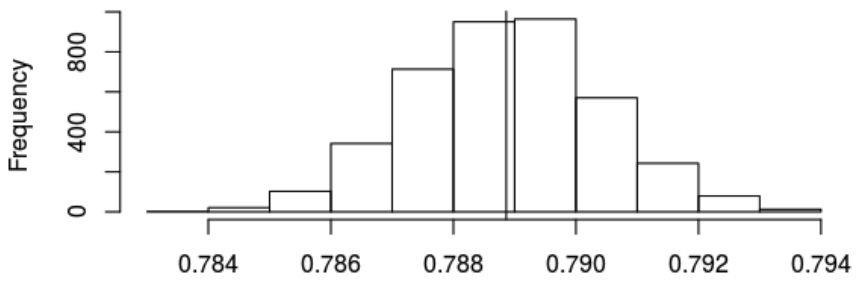

FIGURE 4.2: Posterior classification agreement on ID-defined matches, where matching procedure is performed without IDs.

The overlap is higher among non-matches than matches, indicating that the model tends to agree more with the listed voter registration number non-matches than matches. Given our previously outlined concerns about the state-provided IDs, we expect differences if BRACS correctly identifies the links. However, we may wish to see more overlap among ID-defined matches. Most concerns around the voter registration number are because of a lack of coverage (e.g., the number not being found on the voter file or not being listed for a removed voter) rather than an incorrectly identified match across the files.

4.7 Identifying removed voters using linkage with state-provided IDs

In this section, we link the list of provisional voters to the voter file using the same approach as Chapter 4.6, but we include the voter registration number as a linking variable. To do so, we include an additional comparison vector, γ_{ID}. We set $\gamma_{ij}^{ID} = 1$ if the voter registration number is the same for records A_i and B_j, and $\gamma_{ij}^{ID} = 0$ otherwise.

Including the ID as a linking variable leads to much higher overlap between

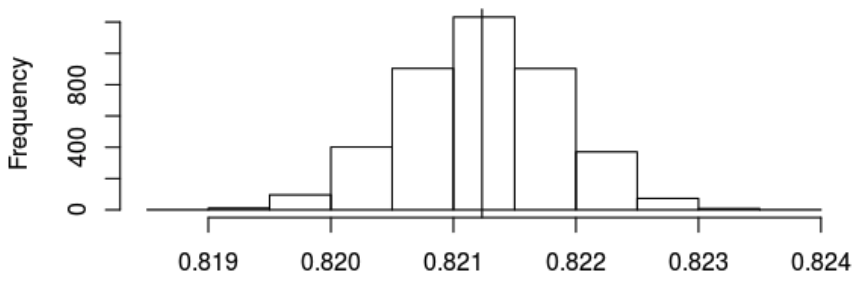

FIGURE 4.3: Posterior classification agreement on ID-defined non-matches, where matching procedure is performed without IDs.

classification agreement for matches, but has less of an impact on the classification agreement for non-matches. As discussed in Chapter 4.6, this difference aligns with our incoming expectations. The state-provided ID can help more confidently match voters who have, for example, changed address or party registration, but a lack of a matching voter registration number could be due in large part to the way that removed voters are classified on the provisional file. For example, if removed voters are tagged as 'Not Registered' since they are not in fact currently registered, this tag may render it less likely that a voter registration number for the removed record is later appended.

We use our linked data set to estimate the number of voters who had been removed from the file before attempting to vote. Our analysis estimates a substantially larger number of removed voters attempt to vote than is reported on the provisional voter list. While using only an ID-based join shows 4,188 provisional voters who had been removed from the voter roll, we estimate that the number of removed voters was closer to 9,333, with a 95% credible interval of (9274, 9392). Both of these numbers are much higher than those tagged as having ballots not counted with a reason of

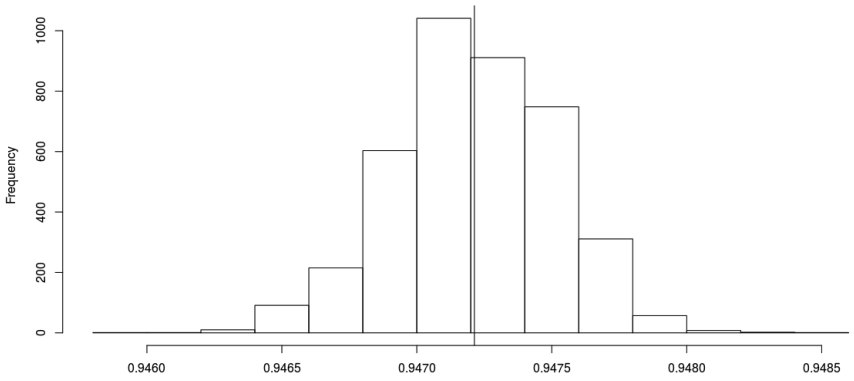

FIGURE 4.4: Posterior classification agreement on ID-defined matches, where matching procedure uses IDs as linking variables.

'Removed' (2,864), though this discrepancy may be explained if the removal status or eligibility status was contested and the vote was eventually counted.

These findings suggest that voter roll purges may have a larger impact than one would see by simply looking at the list of provisional voters or by linking based on the state-provided identifiers. While the voters may not have been eligible for this election, we believe that further research would be beneficial in accurately assessing the impact of purges in states with strict requirements for maintaining an active voter status.

4.8 Conclusions

We do the linkage twice–once ignoring the voter registration number completely and once where we include it as a categorical linking variable. In general, the results using linkage are more similar to each other than to the simple join. Comparing the posterior mode of each Z_i according to our two BRACS analyses, using IDs and not using IDs, the same link for a given provisional voter is identified 91% of the time. Using BRACS and ignoring IDs, we estimate a 78% overlap with the simple ID-based

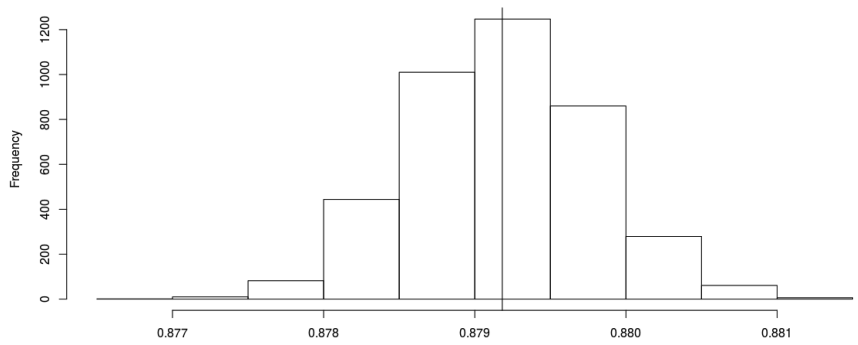

FIGURE 4.5: Posterior classification agreement on ID-defined non-matches, where matching procedure uses IDs as linking variables.

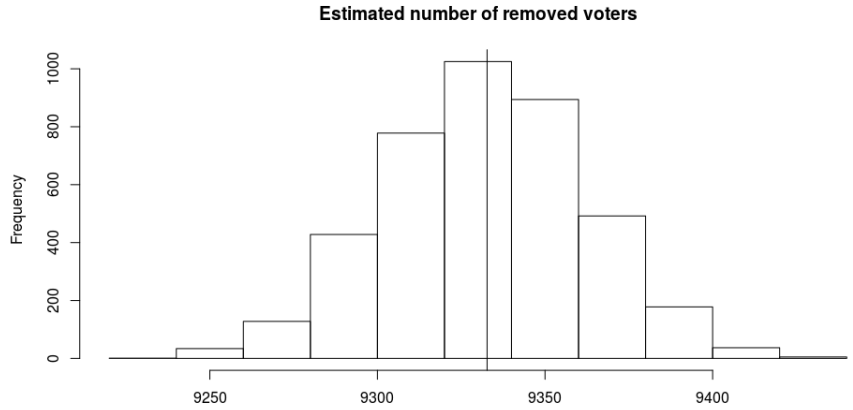

FIGURE 4.6: Posterior distribution of the estimated number of removed voters who cast provisional ballots in the November 8, 2016 election in North Carolina.

join. Using BRACS with IDs as a linking variable, we agree with the ID-based join for 86% of the provisional voters. Despite the discrepancies in the state-provided IDs, we believe that using BRACS with the state-provided IDs leads to more accurate results than ignoring the IDs, especially given our observation that using the IDs impacts the classification agreement of matches more than non-matches.

We show in Table 4.3 that, compared to using state-provided IDs, using BRACS

Table 4.3: Estimates of provisional voters who were removed from the file prior to election day and whose votes were not counted.

	BRACS: no IDs	BRACS: with IDs	IDs only
Removed & not counted	5646 *(5597, 5697)*	5093 *(5043, 5142)*	1247
Removed & listed "Not Registered"	2915 *(2875, 2955)*	2522 *(2481, 2562)*	447

leads to a substantially different estimate of the number of removed voters who attempted to vote on Election Day and did not get their votes counted. I In the table, *Removed and not counted* refers to the number of voters whose provisional votes were not counted and whose most recent voter file status as of Election Day is 'REMOVED'. *Removed and listed "Not Registered"* refers to the number of voters tagged not registered on provisional file but whose most recent voter file status as of Election Day is 'REMOVED'. Linkage approaches use the posterior mean of Z. Intervals are 95% credible intervals. Using BRACS with the IDs as a linking variable, we estimate that 5,093 people who had been removed from the voter file cast provisional ballots which were later rejected. In contrast, there are 2,864 voters on the provisional file who have 'REMOVED' listed as the reason for their vote not counting. Using the ID join shows an even lower estimate of only 1,247 voters whose votes were not counted but whose voter file status according to the Voter Registration Number was 'REMOVED.' This discrepancy may come from a registration in a different county, an error in the voter registration number listed, or a voter registration number listed from an earlier registration on the file. Additionally, many voters appear to have simply been tagged 'Not Registered' on the provisional file, which is also accurate, as they did not have an active registration at the time of voting (see Table 4.3).

We find that analyzing North Carolina's provisional voting file with simple ID-based joins and without probabilistic record linkage may lead to under-estimating the number of people who attempt to vote after being removed from the voter file.

Although the results do not affect the outcome of an election, knowing the registration status can shed light into voting patterns of less consistent voters as well as provide insight into the effects of voter roll purges. If more people are trying to vote after being removed from the file than we have previously expected, then voter roll purges affect turnout more than previously thought. We believe further research using probabilistic record linkage, in North Carolina and other states, is necessary to measure the full impact of purges and active voter status requirements on voter turnout.

5

NC voter linkage with correlated linking variables

5.1 Introduction

In Chapter 4, we showed results from linking the North Carolina 2016 list of provisional voters to the North Carolina voter file. One of the linkage assumptions in the model we used (BRACS) is that the agreement between the linking variables is independent conditional on link status. This assumption implies that, for example, if records A_i and B_j belong to the same person, the fact that both records have the same party registration listed does not give us any information about whether A_i and B_j have the same self-identified race. However, this assumption may not be valid in practice. Race and party registration are highly correlated in North Carolina, and agreement on one may signal agreement on the other, even accounting for link status.

In this chapter, we outline an approach to linking with correlated variables in BRACS. We use simulations to examine how the method compares to a baseline BRACS approach. We then show the results of the election data analysis when accounting for correlation between race and party as linking variables in our analysis of the 2016 NC voter file and provisional voter list.

Table 5.1: Party registration by self-reported race in the linkage comparison space, with cell values corresponding to the row percentage.

	Black	White	Other
Democrat	55.3	35.5	9.3
Republican	4.2	90.8	5.0
Libertarian	10.5	77.1	12.5
Unaffiliated	17.4	67.3	15.2

5.2 Background

Table 5.1 displays the breakdown of voter file party registration by race among the voters in the linkage comparison space. As in Chapter 4, we limit the comparison space to record pairs in the same county and with a normalized Levenshtein similarity of first names and last names above 0.7, as well as a normalized Levenshtein similarity of middle names above 0.25 as potential matches. The distribution of party registration varies by race, and notably between Black and White voters, who make up the majority of the population of the comparison space.

Table 5.2 shows these comparisons of party agreement by race agreement. We see that the two are correlated, particularly in cases where race agrees. For example, if $\gamma_{ij}^{race} = 2$, there is a 71% chance that $\gamma_{ij}^{party} = 2$, contrasted with a 25% chance if $\gamma_{ij}^{race} = 0$. To review the comparison vector coding, if $\text{race}_i = \text{race}_j$, $\gamma_{ij}^{race} = 2$. If $\text{race}_i = \text{B}$ and $\text{race}_j = \text{W}$ or $\text{race}_i = \text{W}$ and $\text{race}_j = \text{B}$, $\gamma_{ij}^{race} = 0$. Otherwise, $\gamma_{ij}^{race} = 1$. Similarly, if $\text{party}_i = \text{party}_j$, $\gamma_{ij}^{party} = 2$. If $\text{party}_i = \text{R}$ and $\text{party}_j = \text{D}$ or $\text{party}_i = \text{D}$ and $\text{party}_j = \text{R}$, $\gamma_{ij}^{party} = 0$. Otherwise, $\gamma_{ij}^{party} = 1$.

Because these comparisons are not also broken out by the unobserved true link status, we cannot say for certain whether the correlation is due to the variables themselves or because of the link status. However, we can compare the distribution to that of two fields that we deem less likely to be correlated. Table 5.2 also shows the comparisons of party agreement by city agreement. The correlations here are

Table 5.2: Comparisons of race agreement and city agreement by party agreement in the linkage comparison space, with cell values corresponding to the row proportion.

		Party agreement: 0	Party agreement: 1	Party agreement: 2
Race agreement	0	0.39	0.36	0.25
	1	0.12	0.41	0.47
	2	0.08	0.21	0.71
City agreement	0	0.18	0.35	0.46
	1	0.11	0.26	0.63

weaker. For example, if $\gamma_{ij}^{city} = 1$, there is a 63% chance that $\gamma_{ij}^{party} = 2$, and a 46% chance if $\gamma_{ij}^{city} = 0$. This difference is less pronounced than it was for race and party.

While we cannot be completely certain of the extent of correlation between variables in the absence of true link status, we believe there is reasonable evidence that comparisons between race and party are correlated. This correlation violates the assumption of conditional independence of comparison vectors. In the next section we will outline the approach we use to address this dependence.

5.3 Methods

Suppose we have two non-independent comparison vectors, γ_1 and γ_2, which compare the first and second comparison fields of A and B. In this application, γ_1 represents γ^{race} and γ_2 represents γ^{party}. γ_1 has L_1 possible values and γ_2 has L_2 possible values. In Chapter 4, we treat γ_1 and γ_2 as independent and therefore multiply their individual likelihoods in the joint likelihood.

We now propose addressing the dependence between γ_1 and γ_2 by concatenating them to form a new linking variable, γ_3, with $L_1 \times L_2$ possible values. For example, if $\gamma_1 = (0, 2, 1, ...)$ and $\gamma_2 = (2, 0, 2, ...)$, then $\gamma_3 = (02, 20, 12, ...)$. We then replace γ_1 and γ_2 in the model with the combined version γ_3.

Table 5.3: Comparison distributions for simulated data in Chapter 5.4.

Field	Γ_f	$\Delta_{ij} = 1$	$\Delta_{ij} = 0$
City	Γ_1	Bern(.8)	Bern(.1)
Gender	Γ_2	Mult(.05, .15, .8)	Mult(.9, .05, .05)
Race	Γ_3	Mult(.05, .15, .8)	Mult(.7, .2, .1)
Party$\|\Gamma_3 = 0$	Γ_4	Mult(.95, .025, .025)	Mult(.7, .2, .1)
Party$\|\Gamma_3 = 1$	Γ_4	Mult(.025, .95, .025)	Mult(.7, .2, .1)
Party$\|\Gamma_3 = 2$	Γ_4	Mult(.05, .05, .9)	Mult(.7, .2, .1)
First name	Γ_5	Mult(.005, .195, .15, .65)	Mult(.95, .02, .02, .01)
Middle name	Γ_6	Mult(.01, .15, .15, .70)	Mult(.95, .02, .02, .01)
Last name	Γ_7	Mult(.05, .05, .20, .70)	Mult(.925, .025, .025, .025)
Address$\|S_1$	Γ_8	Mult(.75, .1, .1, .15)	Mult(.9, 05, .025, .025)
Address$\|S_2$	Γ_8	Mult(.6, .1, .15, .15)	Mult(.8, 15, .025, .025)
Address$\|S_3$	Γ_8	Mult(.5, .1, .2, .2)	Mult(.85, 05, .05, .05)
Address$\|S_4$	Γ_8	Mult(.2, .1, .3, .4)	Mult(.7, .1, .1, .1)

5.4 Simulation studies

As in the simulations in Chapter 3, we imitate possible structure and behavior of the files we link in our application in the following simulations. We use entirely simulated data to evaluate how the proposed method compares to a conditionally independent model.

In order to mimic a collection of two datasets, we generate identification numbers (IDs) A_{i0}: 1-200 for A, representing the list of provisional voters, and B_{j0} 101-600 for the larger B, representing the voter file, such that records with IDs between 101 and 200 are links across the two files. We then generate comparison data for 7 of the 8 comparison fields representing comparisons of city, gender, political party affiliation, first name, middle name, and last name. Our comparison distributions Γ are drawn a according to Table 5.3. In the table, "Field" refers to the comparison field, Γ_f is the comparison vector for that field, and the values in columns $\Delta_{ij} = 1$ and $\Delta_{ij} = 0$ are the distributions of the Γ_f where the pair (A_i, B_j) is and is not a link, respectively.

We use the same hyper-parameters in the simulations and in the application in the

Table 5.4: Values of α_0^k and β_0^k as hyper-parameters for m and u priors.

Field	α_0^k	β_0^k
City	(2, 1)	(1, 20)
Gender	(3, 2, 1)	(1, 5, 10)
First name	(10, 5, 2, 1)	(1, 10, 50, 100)
Last name	(10, 5, 2, 1)	(1, 10, 50, 100)
Middle name	(10, 5, 2, 1)	(1, 10, 50, 100)
Address	(10, 5, 2, 1)	(1, 10, 50, 100)
Race*	(3, 2, 1)	(1, 5, 10)
Party*	(3, 2, 1)	(1, 5, 10)
Race, Party$^+$	(200, 100, 50, 100, 50, 5, 50, 5, 1)	(1, 5, 50, 5, 50, 100, 50, 100, 200)
Race, Party$^-$	(20, 15, 10, 5, 10, 5, 10, 5, 1)	(1, 5, 10, 5, 10, 15, 10, 15, 20)

previous chapters, changing only the prior for the combined race/party comparison. The values of α_0^k and β_0^k are listed in Table 5.4. As mentioned previously, we choose the weights to help guide the model to reasonable posterior modes (rather than all matches or all non-matches) without overpowering the signal in the data.

In running the simulations, we found that, except in cases where the correlation between two variables was very strong, the benefits of accounting for the correlation were often outweighed by the cost of losing the additional prior weight inherent in the extra parameter. However, this risk of losing predictive power can be mitigated by using a strong prior on the combined linking variable. Examples of a weaker and stronger prior are listed in Table 5.4. In the table, * refers to a prior used only in the independent model, and $^+$ refers to the prior used only in the dependent model. $^-$ refers to the weaker prior used in the dependent model in a comparison analysis.

Using the stronger prior, we found that using the correlated structure when two linking variables were not in fact correlated led to similar linkage performance, and performance was improved over the non-correlated structure for both moderate and strong correlations. For additional simulation results, see the appendix.

We perform record linkage using the proposed model and the Gibbs sampler steps and priors outlined in Chapter 3, with variation in the eighth field. For comparison, we use a BRACS model with an assumption of conditional independence between

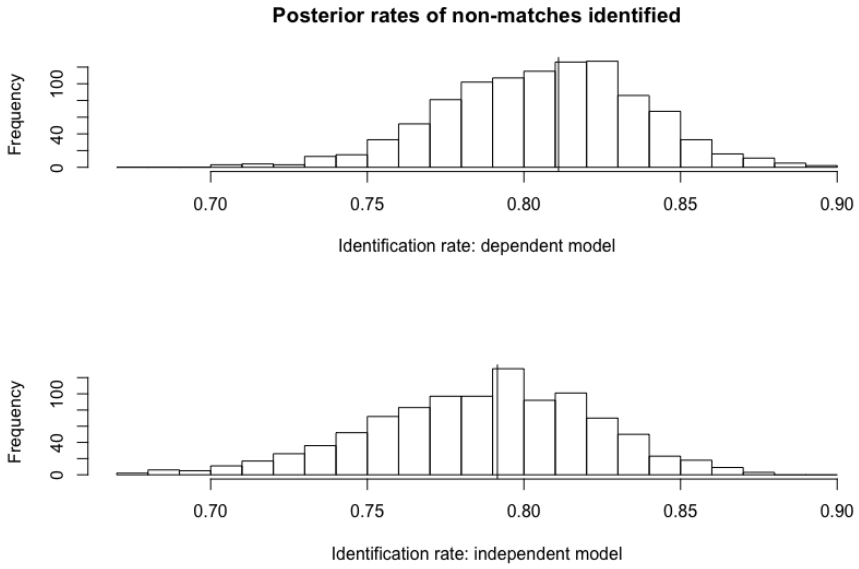

FIGURE 5.1: Posterior rate of non-match identification for the simulation, accounting for and not accounting for field comparison dependence.

γ^{race} and γ^{party}. For both approaches, we generate Levenshtein name similarities for all pairs and consider those with similarities below 0.7 to be non-links. We perform 2,000 iterations with 1,000 burn-in samples. The Levenshtein name similarity cutoff of 0.7 limits the comparison space for computational purposes. It is intended to allow for typographical mistakes, but may exclude some nicknames or name changes.

We compare posterior match identification rates and non-match identification rates for the two approaches. We define match identification rates as the proportion of the 100 true matches in A that are correctly identified in each post burn-in draw of Z, or $\frac{1}{100}\sum_{i=101}^{200}\mathbb{1}(Z_i = A_{i0})$ and non-match identification rates as the proportion of the 100 non-matches in file A that are correctly classified as non-matches, or $\frac{1}{100}\sum_{i=1}^{100}\mathbb{1}\big(Z_i = (N_B + 1)\big)$.

The match identification rates are nearly 100% for both approaches, but we see differences in the non-match identification rate, as shown in Figure 5.1.

Our simulation studies suggest that we can improve accuracy if we account for

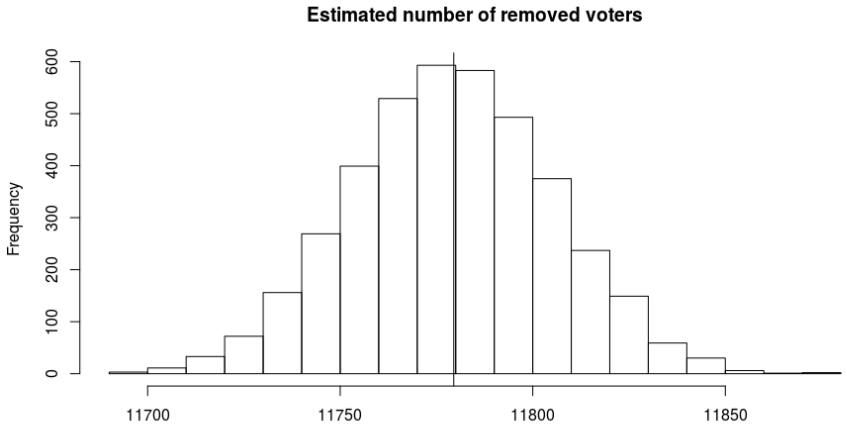

FIGURE 5.2: Estimated number of removed voters who cast provisional ballots in the November 2016 election in NC, with matching procedure outlined in Chapter 5.5.

dependent comparison fields. However, the choice of prior can impact the analysis. A weak prior can lower non-match identification rates in particular, and so using a relatively strong prior can compensate for the decrease in the number of parameters. It is worth noting, however, that in scenarios where both the match identification rates and the non-match identification rates are already high, neither the choice of priors nor the accounting for dependence have much effect on the results.

5.5 Linking the NC provisional voter list to the voter file: analysis and discussion

When accounting for linking variable correlation in the final analysis, we count a higher number of removed voters in the subsequent analysis on the linked data, as shown in Figure 5.2. However, we see similar overlap rates with the state-provided IDs as we did with the previous linkage, as shown in Figures 5.3 and 5.4.

We also performed the analysis with a weaker prior distribution on the combined field and the differences in the resulting analysis were minimal. See Figure 5.5 for the posterior distribution of the estimated number of removed voters. We find the

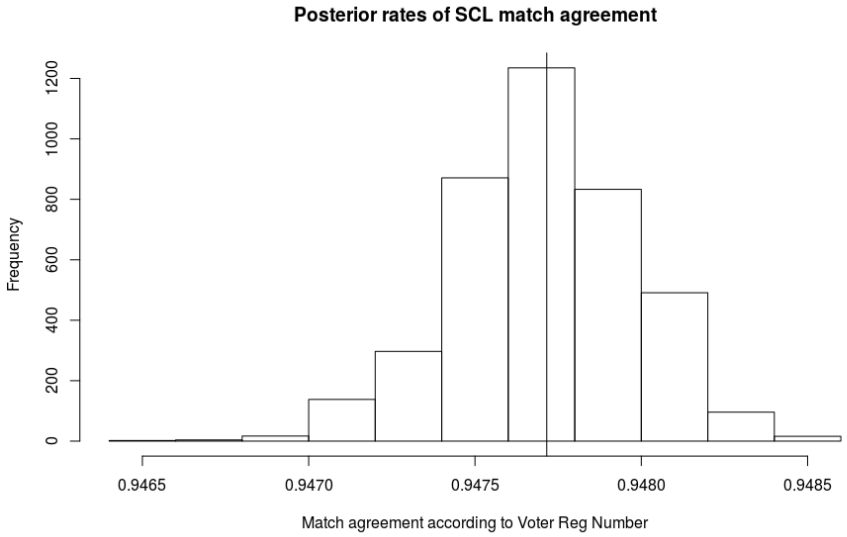

FIGURE 5.3: Posterior classification agreement on ID-defined matches, with matching procedure outlined in Chapter 5.5.

similarities of these results encouraging, because they follow patterns we observed in the simulations, where scenarios with high separation between the linked and non-linked distributions resulted in strong enough performance that changing the priors or even aspects of the model specification had little impact on the results.

5.6 Conclusions

In conclusion, we find in our analysis that incorrectly assuming independence of comparison fields conditional on link status may lead to degradation in link classification accuracy. Using a combined linking field can improve the accuracy, particularly in cases when the correlation between the comparison vectors is high.

We use this combined-variable approach, incorporated in the BRACS model, to estimate the number of voters in North Carolina who cast provisional ballots in 2016 and had been removed from the voter file. Similar to the previous chapter, our analysis suggests that using only the state-provided identifiers or the labels provided on the provisional file may lead to an under-estimation of the number of voters

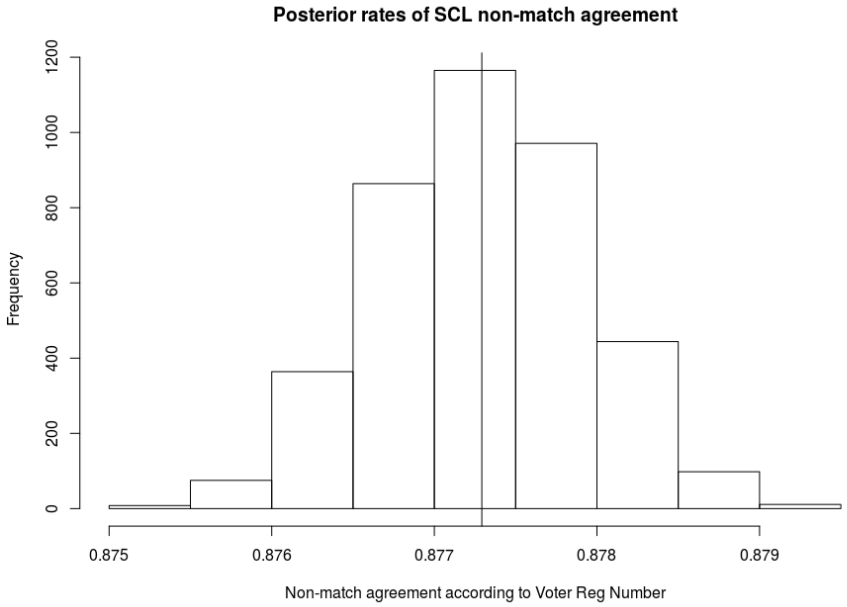

FIGURE 5.4: Posterior classification agreement on ID-defined non-matches, with matching procedure outlined in Chapter 5.5.

attempting to vote on Election Day. While this under-estimation would not affect election outcomes, it is useful information for understanding the effect of voter roll purges. If more people are trying to vote after being removed from the file than we have previously expected, then voter roll purges affect turnout more than previously thought.

While we believe that BRACS, particularly accounting for inter-field dependence, is a good methodological approach for this application, we acknowledge the limitations inherent in record linkage. Even a strong algorithm can mis-classify records, and so we cannot be completely confident in analysis derived from the procedure. However, we believe that using advanced record linkage techniques can help shed light on potential issues with existing records, and how those record-keeping issues can affect the estimated implications of voter roll purges.

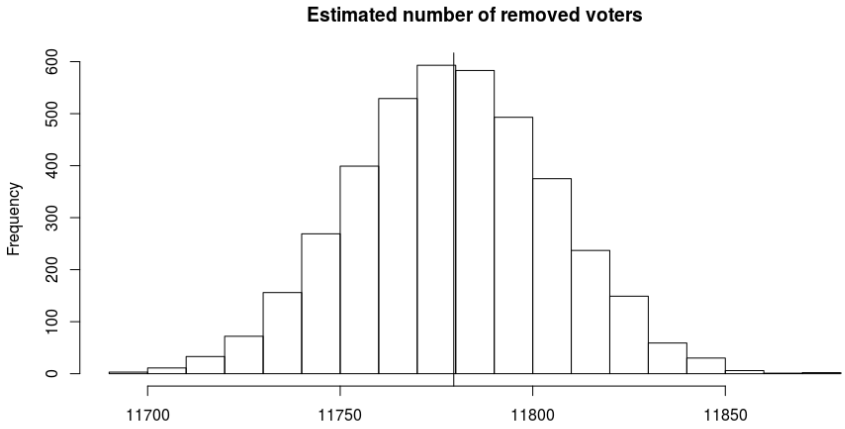

FIGURE 5.5: Estimated number of removed voters who cast provisional ballots in the November 2016 election in NC with matching procedure outlined in Chapter 5.5 and a weaker prior on the combined field, shown in Table 5.4.

6

Conclusions

6.1 Chapter 2

Methods for causal inference and record linkage have developed independently, but the simulation results indicate that it can be fruitful to consider methods that explicitly account for both tasks. For settings where covariates and assignments are in one file and outcomes are in the other file, our simulations suggest that case selection strategies can improve causal estimates for analyses based on propensity score subclassification. In these simulations, arguably ETSR performs best overall. It has the smallest mean squared errors in some scenarios, and when other case selection rules have lower mean squared errors, ETSR generally is not far behind. Of course, as with any simulation study, these findings are based on limited simulation studies and particular assumptions. Additional research is needed to assess the performance of the case selection strategies when these assumptions are not reasonable.

As with any methodology, there are scenarios where the case selection procedures may not be effective. In particular, when the known links include outliers that pull the estimate of $\hat{\tau}_{\mathcal{L}_0}$ far away from τ, the procedures might not add many cases, even

true links, to the data used in the causal estimate, as doing so could cause the estimated variance to increase. Additionally, the case selection procedures may give misleading results when their underlying assumptions, in particular independence of the linking variables and constant treatment effects, are unreasonable. Finally, the procedures may suffer when the underlying analysis models are poorly specified, including the propensity score models, subclass boundaries, and regressions for adjusted inferences.

We recommend being sensible in the choice of linkage technology. We found that adding many incorrect links, e.g., by using very low thresholds to accept almost any proposed link, can reduce the estimated variance of the treatment effect due to the increased sample size. However, this results in poor quality estimates of τ. When reasonable cutoffs on linkage scores are enforced, the pattern of the estimated variance under varying thresholds tends to be U-shaped. However, the estimated variance can become S-shaped when large numbers of incorrect links are added. Using the ETSR limits the possibility of favoring thresholds corresponding to high numbers of incorrect links, but we still emphasize the importance of using sound record linkage techniques when making causal inferences with linked data.

6.2 Chapter 3

The results from the BRACS simulation studies suggest that clustered linkage can improve linkage quality when Γ differs by group. In both simulation scenarios, we saw the most improvement in the identification of non-matches. When the distribution of an agreement field, such as address, varies in a systematic way, the m and u parameters will not reflect the agreement probabilities appropriately. By allowing the m and u parameters to change by group, we allow for more precise estimation of the likelihood of each pair being a match, and thus higher classification rates.

When Γ does not differ by group, adding dependency on S_k may add noise in

the estimates due to smaller sample sizes available to estimate the parameters m_k and u_k. It may also make linkage more sensitive to a strong prior. These strengthened assumptions can help accuracy, as there are certain contexts (including voter file linkage) in which there has been substantial work previously done, and prior information may be both useful and easily obtained.

There can be increased variance in using BRACS, but there can be substantial gains in accuracy, particularly when there is large variation in comparison distributions conditional on k, or when matches or non-matches are difficult to identify. In addition to the gains in accuracy, the Bayesian nature of the model also allows us to account for linkage uncertainty, which can then be used in further analytic decisions. These may include further modeling of voter behavior, or simply voter identification for follow-up. The posterior distribution provided by this approach allows for this post-linkage flexibility.

There are challenges inherent in the approach as well. Adding additional parameters increases the computational complexity, which is already a challenge for Bayesian record linkage approaches. We mitigate this problem by imposing additional restrictions to our comparison space based on name, but depending on context, the additional complexity due to the additional parameters may not be worth the computational cost. Additionally, there is a challenge in deciding which fields to condition on. In some cases, there may be built-in categorizations (like age group on the voter file) and contextual knowledge to guide the decision. However, the complexity increases non-linearly with the number of conditional fields. How to account for higher numbers of conditional fields is a topic for future research.

6.3 Chapter 4

We find that analyzing North Carolina's provisional voting file with simple ID-based joins and without probabilistic record linkage may lead to under-estimating the num-

ber of people who attempt to vote after being removed from the voter file. Although the results do not affect the outcome of an election, knowing the registration status can shed light into voting patterns of less consistent voters as well as provide insight into the effects of voter roll purges. If more people are trying to vote after being removed from the file than we have previously expected, then voter roll purges affect turnout more than previously thought. We believe further research using probabilistic record linkage, in North Carolina and other states, is necessary to measure the full impact of purges and active voter status requirements on voter turnout.

6.4 Chapter 5

We find in our analysis that incorrectly assuming independence of comparison fields conditional on link status may lead to degradation in link classification accuracy. Using a combined linking field can improve the accuracy, particularly in cases when the correlation between the comparison vectors is high.

We use this combined-variable approach, incorporated in the BRACS model, to estimate the number of voters in North Carolina who cast provisional ballots in 2016 and had been removed from the voter file. Similar to Chapter 4, our analysis suggests that using only the state-provided identifiers or the labels provided on the provisional file may lead to an under-estimation of the number of voters attempting to vote on Election Day. While this under-estimation would not affect election outcomes, it is useful information for understanding the effect of voter roll purges. If more people are trying to vote after being removed from the file than we have previously expected, then voter roll purges affect turnout more than previously thought.

While we believe that BRACS, particularly accounting for inter-field dependence, is a good methodological approach for this application, we acknowledge the limitations inherent in record linkage. Even a strong algorithm can mis-classify records, and so we cannot be completely confident in analysis derived from the procedure.

However, we believe that using advanced record linkage techniques can help shed light on potential issues with existing records, and how those record-keeping issues can affect the estimated implications of voter roll purges.

7

Appendix

7.1 Online Supplement for Chapter 2

7.1.1 Introduction

In this supplement, we present additional simulation results, organized as follows. In Chapter 7.1.2, we show results from simulation studies using the same data generation methods as the main text, but using a regression correction within subclasses as part of the causal analysis. In Chapter 7.1.3, we examine the simulation results from the main text simulations using the ETSR with values of k ranging from 0.1 to 3. In Chapter 7.1.4, we present simulation results for two additional scenarios: one using a smaller size of File B and one using average Jaro-Winkler scores as the record linkage technique. We also present results for three scenarios where some of the assumptions used in deriving the methods are not true. Specifically, we run simulations where the treatment effect is not constant but linkage errors are still independent of all variables, where the treatment effect is not constant and linkage errors are correlated with a causally relevant variable, and where one of the linking variables is a confounder in the causal analysis.

7.1.2 Original simulation studies with regression correction

Here we present results of simulation studies evaluating the performance of the case selection algorithms outlined in Chapter 2.4 with an additional within-subclass regression correction in the causal analysis. We use the same data generation process as in Chapter 2.5.1.

Within each subclass, we use a regression adjustment to estimate τ. Specifically, for any $\mathcal{L}_h$, within each subclass j we estimate the model, $y^* = \alpha_{0j} + x_1^*\alpha_{1j} + x_2^*\alpha_{2j} + w^*\beta_j + \epsilon_j$, where $\epsilon_j \sim N(0, \sigma^2)$. We estimate this linear model for each simulation run, including the scenario with a non-linear response surface. We estimate all parameters via ordinary least squares. To estimate τ, we use $\hat{\tau}_\beta^* = \sum_{j=1}^{J} \lambda_j \hat{\beta}_j^*$, where each $\hat{\beta}_j^*$ is the estimate of β_j. We estimate the variance using (10) in Chapter 2.3, replacing the two-sample variance estimator with the estimated variance of $\hat{\beta}_j^*$ from each within-subclass regression.

To provide context for the linkage quality, we repeat Table 2.1 from the main text. As noted in the main text, the quality of links at thresholds of 9.8 and above is high but deteriorates quickly as one drops the threshold, with more false links and duplicates. Table 7.2 shows examples of the linked data under different thresholds, illustrating the types of errors tolerated when decreasing the threshold. Field values are shown for first name, last name, birth day, birth month, and birth year in files A and B, along with true (but unobserved) link status.

Figure 7.1 summarizes the distributions of $\hat{\tau}_\beta^*$ and its estimated variance for 100 independent runs of the first simulation with a linear response surface and $(\tau = 50, \sigma = 10)$ at all qualifying values of the threshold for selecting cases, illustrating the behavior of the estimates under various thresholds. Similar to analysis without a regression correction, the choice of threshold matters for the quality of the causal estimate. Using thresholds below 9.8 includes incorrect links that degrade the

Table 7.1: Linkage summary under various thresholds, repeated from main text.

Threshold	Link Rate	Units	Duplicates
0.3	76.2 %	1309	27
0.8	84.6 %	1179	8
1.6	91.6 %	1088	3
2.1	95 %	1048	0
2.8	97.5 %	1018	0
9.3	98.7 %	1002	0
9.8	99.2 %	973	0
10.5	99.7 %	775	0
11.8	100 %	638	0
19.5	100 %	531	0

Table 7.2: Examples of new links added under various thresholds.

Thresh.	F. Name:A	F. Name:B	L. Name:A	L. Name:B	M:A	D:A	Y:A	M:B	D:B	Y:B	Status
0.30	FRIEDA	GERHARD	MUELLER	MUELLER	8	25	1941	8	14	1941	False
0.80	RENATE	RENATE	SCHMIDT	WERNER	11	12	1939	11	19	1939	False
1.61	PETRA	KLAUS	SCHMITT	SCHMITT	7	14	1958	6	14	1958	False
2.10	KLAUS	KLAUS	WAGNER	KUEHN	5	14	1968	9	14	1968	False
2.83	HEINZ	HEINZ	MAYER	MAYER	7	13	1949	12	2	1949	False
9.26	PAUL	PAFUL	PFEIFFER	PFEIFFER	10	20	1956	10	20	1956	True
9.76	CHRISTINE	CHRISTINE	MUELLER	MUEKLER	7	18	1937	7	18	1937	True
10.49	BAERBEL	BAERBEL	FISCHER	FISCHER	8	7	1976	8	4	1976	True
11.80	FRANK	FRANK	PETERS	PETERS	6	1	1990	7	1	1990	True
19.45	GERTRUD	GERTRUCD	MUELLER	MUELLER	11	19	1986	11	19	1986	True

accuracy of $\hat{\tau}^*_\beta$. On the other hand, using the highest threshold values cause $\hat{\tau}^*_\beta$ to be based on relatively small numbers of individuals, which results in relatively high variances of $\hat{\tau}^*_\beta$.

Table 7.3 displays key results of the simulation runs. First examining the simulation setting with $(\tau = 50, \sigma = 10)$, all three stopping rules reduce mean squared errors compared to using only known links. All of the case selection methods have increased mean squared errors compared to using the true links, reflecting the information loss from having to use inexact linkage. In this simulation, MEV offers the most substantial reductions in mean squared error, although all three methods have comparable performances. The next two simulations with $\tau = 10$ and $\tau = 1$ have qualitatively similar results.

In the fourth scenario where we return $\tau = 50$ and increase the variance to $\sigma = 25$, we again see that the case selection methods have larger mean squared errors than

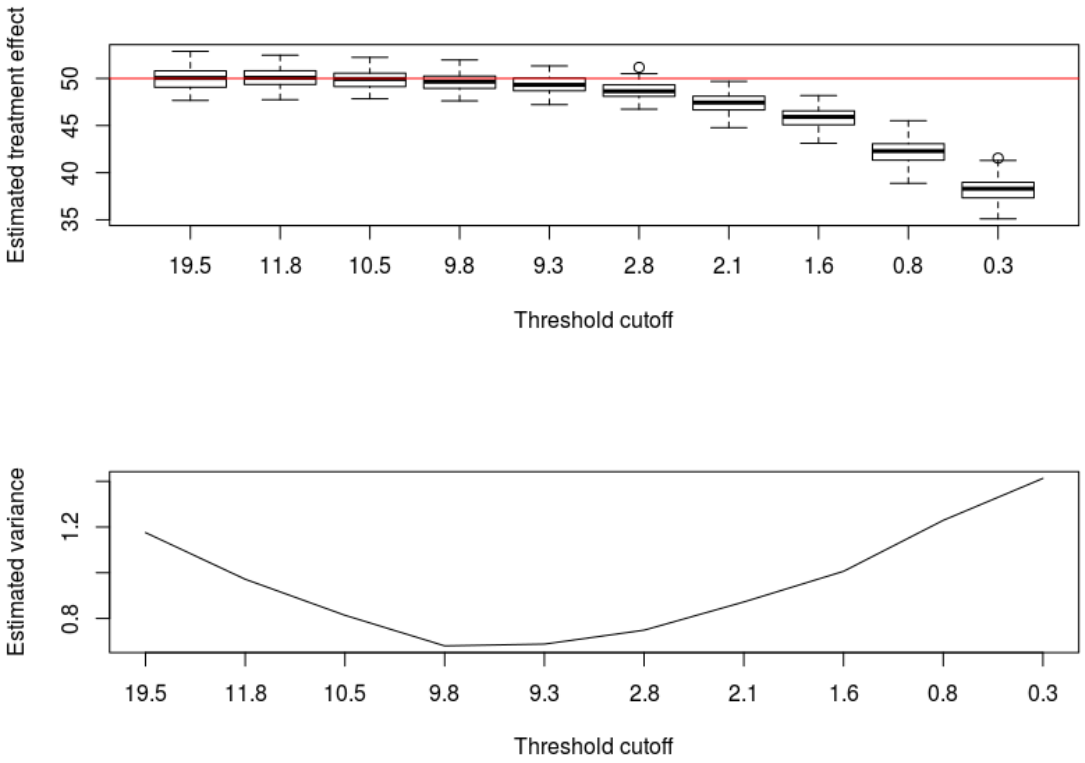

FIGURE 7.1: Distribution of 100 point and variance estimates of regression-adjusted treatment effects for simulation with Fellegi-Sunter linkage and constant treatment effect, where $\tau = 50$ and $\sigma = 10$.

using the true links, as one would expect generally. Here, however, the MEV does not perform as well as using only the known links. With MEV, the estimated variance of $\hat{\tau}^*_\beta$ is substantially smaller than its true variance, which happens because the MEV criterion accepts too many false links, as evidenced by the average threshold level of 3.5. Because of the large variance in the outcomes, accepting false links tends to have greater impact on the bias and mean squared error of $\hat{\tau}^*_\beta$. As a result, MEV increases bias and decreases the accuracy of the variance estimation. This illustrates the concerns about MEV noted at the beginning of Chapter 2.4.2 and used to motivate ETSR. In contrast, ETSR and MEDOV continue to outperform using the known links only, with MEDOV offering slightly larger reductions than

Table 7.3: Summary of results across 100 runs of the six simulation scenarios with Fellegi-Sunter linkage and a regression correction within subclasses.

Scenario	Linkage	Mean $\hat{\tau}^*_\beta$	Var $(\hat{\tau}^*_\beta)$	$\hat{var}(\hat{\tau}^*_\beta)$	MSE $\hat{\tau}^*_\beta$	Avg. h
Linear, $(\tau = 50, \sigma = 10)$	Perfect	50.0	0.6	0.1	0.6	
	Known	49.9	4.9	4.5	4.9	
	MEV	49.6	0.9	0.7	1.1	9.6
	ETSR	49.9	1.4	0.8	1.3	10.8
	MEDOV	49.8	1.2	0.9	1.2	13.1
Linear, $(\tau = 10, \sigma = 10)$	Perfect	10.1	0.7	0.1	0.7	
	Known	10.2	5.7	4.4	5.7	
	MEV	10.0	0.8	0.6	0.8	9.3
	ETSR	10.1	1.3	0.8	1.2	9.2
	MEDOV	9.7	1.2	0.9	1.3	11.1
Linear, $(\tau = 1, \sigma = 10)$	Perfect	1.1	0.7	0.1	0.7	
	Known	1.2	5.7	4.4	5.7	
	MEV	1.1	0.8	0.6	0.8	9.3
	ETSR	1.1	1.2	0.8	1.2	10.2
	MEDOV	0.9	1.1	0.9	1.1	12.7
Linear, $(\tau = 50, \sigma = 25)$	Perfect	50.2	4.3	0.8	4.3	
	Known	50.6	35.4	27.7	35.5	
	MEV	44.3	32.0	3.8	63.8	3.5
	ETSR	50.0	7.7	4.4	7.6	8.7
	MEDOV	49.9	7.1	5.6	7.1	13.8
Linear, High R^2, $(\tau = 50, \sigma = 10)$	Perfect	50.1	0.7	0.1	0.7	
	Known	50.2	5.5	4.4	5.5	
	MEV	49.8	0.9	0.8	1.0	9.9
	ETSR	50.1	1.3	0.9	1.3	11.7
	MEDOV	47.8	19.8	1.6	24.5	9.8
Non-linear, $(\tau = 50, \sigma = 10)$	Perfect	50.9	16.9	3.5	17.6	
	Known	51.0	98.1	79.7	98.2	
	MEV	45.5	31.7	7.1	51.9	5.5
	ETSR	49.7	15.7	9.3	15.7	9.5
	MEDOV	47.9	25.4	7.6	29.6	9.8

ETSR.

In the fifth scenario, MEV has the smallest mean squared error among the case-selection procedures, though ETSR is not far behind. In contrast, MEDOV has the worst performance, driven by both the higher bias and the higher variance. MEDOV results in large variation in the selected thresholds. The simulated standard error of the threshold choice for MEDOV across the 100 runs of this simulation is around 6.6, compared to 0.7 for MEV.

In the sixth scenario where the response surface is non-linear, all three case selection procedures are again preferable to using known links alone. Here ETSR offers the greatest reductions in mean squared error.

7.1.3 ETSR tether parameter

In this section we present results from running the simulation scenarios in Chapter 2.5 of the main text using the ETSR with different values of the tether parameter k. We use the same methods of analysis as the main text: Fellegi-Sunter record linkage and subclassification on propensity scores without a regression correction.

As evident in Table 7.4, overall the results with k between 0.5 and 2 are qualitatively similar. Sometimes, however, setting $k = 0.1$ results in too conservative of a threshold choice, and setting $k = 3$ results in too liberal of a threshold choice. Although $k = 0.5$ is a reasonable compromise, analysts may want to choose k based on the context of the analysis and the relative consequences of including too many false links or not enough true links. In the table, k refers to the number of standard errors used in the tether restriction. $\text{Var}(\hat{\tau}^*)$ refers to the empirical variance of the estimated treatment effects across each set of 100 runs, and $\hat{var}(\hat{\tau}^*)$ refers to the average of the estimated variances across each set of 100 runs. Avg. h refers to the average threshold chosen across the 100 runs.

7.1.4 Additional simulation studies

Simulation with smaller size for File B

In the first simulation, we examine the performance of the algorithm when the two files to be linked are the same sizes. Instead of linking File A to a File B with 8,000 records, we leave only 2,000 records in File B, half of which have a link in File A. We use the data generation process and record linkage techniques in the first scenario in Chapter 2.5, i.e., the response surface is linear with $(\tau = 50, \sigma = 10)$, and generate 100 independent simulation runs.

Table 7.5 displays the link rate, number of linked units, and the number of units in File B used multiple times as links for the different possible thresholds. The link rates here are higher than in the other simulations, since there are fewer false links

Table 7.4: Summary of results across 100 runs of the six simulation scenarios with Fellegi-Sunter linkage and propensity score subclassification.

Scenario	k	Mean $\hat{\tau}$*	Var ($\hat{\tau}$*)	$\hat{var}(\hat{\tau}$*)	MSE $\hat{\tau}$*	Avg. h
Linear, ($\tau = 50, \sigma = 10$)	0.1	47.1	2.1	1.6	10.7	
	0.5	47.0	1.9	1.2	11.0	
	1	46.7	1.2	1.1	11.8	9.3
	2	46.6	1.1	1.0	12.6	9.1
	3	46.6	1.1	1.0	12.7	9.1
Linear, ($\tau = 10, \sigma = 10$)	0.1	7.1	2.1	1.6	10.5	
	0.5	7.1	1.7	1.1	9.9	
	1	7.1	1.2	1.0	9.7	7.9
	2	7.0	1.1	1.0	9.9	7.5
	3	7.0	1.1	1.0	9.9	7.5
Linear, ($\tau = 1, \sigma = 10$)	0.1	-1.9	2.0	1.5	10.5	
	0.5	-1.8	1.7	1.2	9.8	
	1	-1.8	1.2	1.0	9.2	7.6
	2	-1.8	1.1	1.0	8.9	7.6
	3	-1.8	1.1	1.0	8.9	7.6
Linear, ($\tau = 50, \sigma = 25$)	0.1	47.2	9.4	6.3	17.0	
	0.5	47.0	8.1	4.8	17.1	
	1	46.7	6.3	4.3	17.0	7.7
	2	46.4	5.9	4.1	19.0	7.1
	3	46.2	7.7	4.1	22.3	6.9
Linear, High R^2, ($\tau = 50, \sigma = 10$)	0.1	44.0	14.2	10.7	50.4	
	0.5	43.8	12.9	8.1	51.5	
	1	43.5	10.6	7.4	53.1	7.1
	2	43.1	8.3	7.1	56.3	6.5
	3	42.6	12.5	7.1	66.9	6.2
Non-linear, ($\tau = 50, \sigma = 10$)	0.1	47.7	21.9	14.6	26.9	
	0.5	47.0	17.8	9.5	26.7	
	1	46.0	15.2	6.5	30.9	7.9
	2	44.6	20.4	6.0	49.4	6.2
	3	43.8	26.1	5.8	64.6	5.9

possible with the smaller size of File B. Therefore, even at the lowest threshold above zero, we still see a link rate of 97.9 %. Table 7.6 shows randomly selected examples of links added under the various thresholds. Although the thresholds are numerically similar to the other simulations, the implications on linkage quality are clear. The pool of possible links is of higher quality due to the smaller size of File B, so the links added are more likely to correspond to the same person.

Figure 7.2 and the top panel of Table 7.7 summarize the results. All the case selection procedures offer improvements over using the known links alone, with the best performance using ETSR.

Table 7.5: Linkage summary under various thresholds for simulation with smaller File B size.

Threshold	Link Rate	Units	Duplicates
0.6	97.9 %	1020	0
0.8	98.7 %	1011	0
1.3	99.8 %	996	0
8	99.9 %	991	0
8.3	99.9 %	967	0
8.8	100 %	773	0
9.8	100 %	638	0
17.3	100 %	531	0

Table 7.6: Examples of new links added under various thresholds for simulation with smaller File B size.

Thresh.	F. Name:A	F. Name:B	L. Name:A	L. Name:B	M:A	D:A	Y:A	M:B	D:B	Y:B	Status
0.57	JUERGEN	ANGELIKA	SCHULZ	SCHULZ	7	24	1947	6	24	1947	False
0.82	WALTER	WALTER	KOEHLER	MEYER	7	18	1935	9	18	1935	False
1.33	ROBERT	ROBERT	LANG	LANG	9	24	2007	2	22	2007	True
8.05	ELDKE	ELKE	WEISS	WEISS	4	30	1978	4	30	1978	True
8.30	KARIN	KARIN	MUELLRR	MUELLER	2	9	1974	2	9	1974	True
8.80	RENATE	RENATE	HORN	HORN	11	18	1994	11	81	1994	True
9.85	RUTH	RUTH	MEIER	MEIER	11	29	1961	1	29	1961	True
17.32	STEFEAN	STEFAN	STEIN	STEIN	11	21	1938	11	21	1938	True

Simulation with Jaro-Winkler scores

The second set of simulations uses average Jaro-Winkler scores for record linkage and a constant treatment effect. We generate the data in the same way as the first simulation in Chapter 2.5 of the main text, but change the record linkage method.

We use the similarity metric $S(\gamma_{ii'}) = (1/4)\sum_{f=1}^{4}\Phi_{JW}(fii')$, where $\Phi_{JW}(fii')$ is the Jaro-Winkler similarity of the comparison field f for record pair (i, i'). We require $S(\gamma_{ii'}) \geqslant 0.8$ for the pair to be considered a possible link. As evident in Table 7.8, the quality of links at thresholds of 0.9 and above is high, with a link rate upwards of 99%, but it deteriorates at a threshold of 0.8. Table 7.9 shows examples of new links under various thresholds.

Figure 7.3 summarizes the distributions of $\hat{\tau}^*$ and its estimated variance for 100 independent runs of the simulation that uses the linear outcome distribution described for the first simulation in Chapter 2.5 with $\tau = 50, \sigma = 10$). Once again, the choice of threshold matters for the quality of the causal estimate, although all

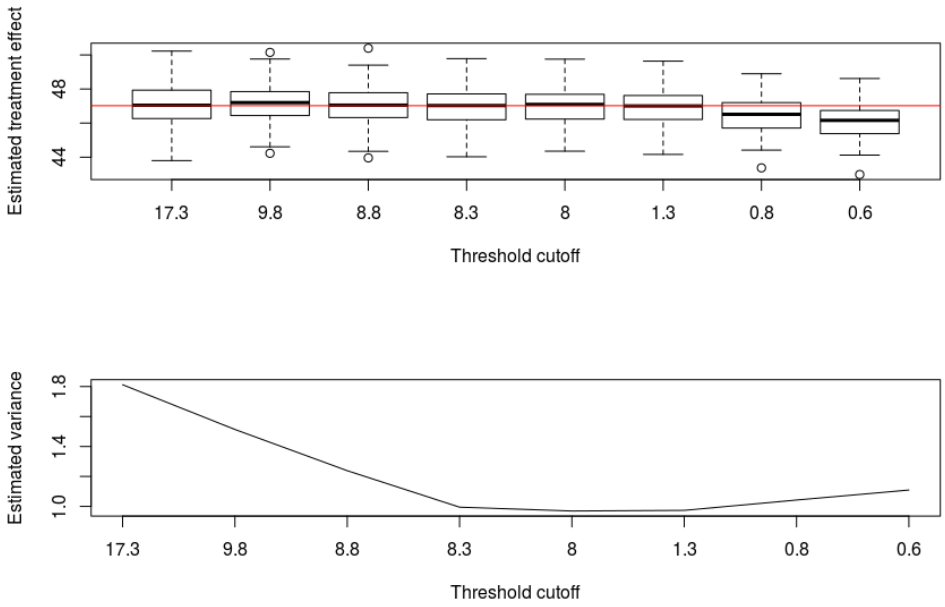

FIGURE 7.2: Distribution of 100 point and variance estimates of treatment effects for simulation with Fellegi-Sunter linkage and constant treatment effect with smaller File B.

are reasonably high quality. The sweet spot reflecting a close-to-optimal trade off in contributions to mean-squared error is a threshold somewhere around 0.9, which provides point estimates clustered most closely around the value of $\hat{\tau}$ attainable with perfect record linkage. As is evident in the bottom panel of Figure 7.3, across the 100 runs the estimate of the variance tends to be minimized when the threshold is around 0.9.

Table 7.7 summarizes results of treatment effect estimation when using the different rules for selecting links. Compared to using only the known links with all fields having $\gamma_{fii'} = 1$, using any of the proposed threshold rules reduces the mean squared error of $\hat{\tau}^*$ to the point where results are similar to those based on the true links. MEV and ETSR tend to result in smaller mean square errors than MEDOV, although the difference is minor.

Table 7.7: Summary of results across 100 runs of the additional simulation scenarios with propensity score subclassification using $\hat{\tau}^*$ as the estimator of treatment effect.

Scenario	Linkage	Mean $\hat{\tau}^*$	Var ($\hat{\tau}^*$)	$\hat{var}(\hat{\tau}^*)$	MSE $\hat{\tau}^*$	Avg. h
$n_B = 2000, (\tau = 50, \sigma = 10)$	Perfect	47.0	1.0	0.9	9.9	
	Known	46.9	6.2	6.3	16.1	
	MEV	46.9	1.0	1.0	10.4	3.9
	ETSR	**47.0**	**1.3**	**1.1**	**10.0**	**5.8**
	MEDOV	46.9	1.2	1.4	11.1	10.3
Jaro-Winkler linkage, $(\tau = 50, \sigma = 10)$	Perfect	47.1	1.1	0.9	9.3	
	Known	47.3	7.2	6.5	14.3	
	MEV	**47.0**	**1.2**	**1.1**	**10.4**	**0.9**
	ETSR	47.1	2.1	1.3	10.6	0.9
	MEDOV	47.0	2.0	1.5	11.2	1.0
Non-constant treatment effect, $(\tau = 50, \sigma = 10)$	Perfect	46.3	2.2	1.8	15.7	
	Known	46.2	13.8	13.2	28.4	
	MEV	45.9	2.5	1.9	19.2	9.0
	ETSR	**46.2**	**2.7**	**2.2**	**17.2**	**9.8**
	MEDOV	46.0	2.7	2.8	18.9	13.3
Linkage quality correlated with treatment effect, $(\tau = 50, \sigma = 10)$	Perfect	46.3	1.4	1.8	15.2	
	Known	57.1	11.6	11.5	62.4	
	MEV	**46.1**	**1.7**	**1.8**	**16.7**	**8.9**
	ETSR	57.0	3.1	2.9	52.1	19.4
	MEDOV	51.5	33.5	2.4	35.3	14.1
Linking variable correlated with outcome, $(\tau = 50, \sigma = 10)$	Perfect	47.8	1.8	3.2	6.5	
	Known	48.7	19.6	23.7	21.2	
	MEV	46.8	5.6	3.2	15.9	7.4
	ETSR	**47.7**	**3.5**	**3.8**	**8.5**	**9.2**
	MEDOV	47.6	2.5	4.5	8.5	12.6

Table 7.8: Linkage summary under various thresholds for Jaro-Winkler linkage simulation.

Threshold	Link Rate	Units	Duplicates
0.8	88.5 %	956	0
0.9	99.5 %	850	0
0.9	99.8 %	834	0
1	100 %	712	0
1	100 %	612	0
1	100 %	493	0

Simulation with a non-constant treatment effect

The case selection algorithms are developed assuming constant treatment effects. In some situations, treatment effects may differ across individuals. In this section, we examine the performance of the algorithm when the treatment effect is non-constant. The data generation uses the same approach as in Chapter 2.5.1, but we do not set $\tau_i = 50$ for all units. Instead, for each individual i, we generate an additional binary

Table 7.9: Examples of new links added under various thresholds.

Thresh.	F. Name:A	F. Name:B	L. Name:A	L. Name:B	M:A	D:A	Y:A	M:B	D:B	Y:B	St.
0.80	KARIN	ULRIKE	SCHNEIDER	SCHNEIDER	9	12	1954	9	22	1954	F
0.90	SABINE	SABINE	FRANK	FRANK	4	22	1931	4	72	1931	T
0.92	STEFAN	STEFAN	WAGNER	WAGNER	8	21	1926	8	22	1926	T
0.98	JUERGEN	JUERGEN	MURLLER	MUELLER	4	14	2002	4	14	2002	T
0.98	WOLFGANG	WOLFGANG	FISCHWR	FISCHER	2	26	1967	2	26	1967	T
0.99	NORBERT	NORBERT	KAISER	KAISER	4	11	1934	4	11	1934	T

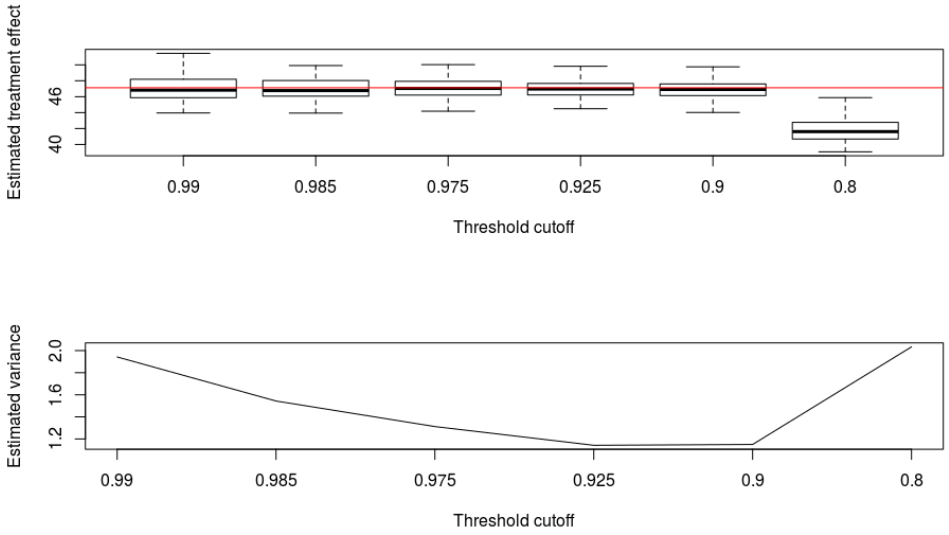

FIGURE 7.3: Distribution of 100 point and variance estimates of treatment effects for simulation with Jaro-Winkler score linkage and constant treatment effect.

variable x_{i3} from a Bernoulli distribution with probability equal to $(0.4w_i + .6(1-w_i))$, and set $\tau_i = 50 + 20x_{i3} - 20(1 - x_{i3})$. Hence, treatment effects are dependent on a binary covariate that is not equally distributed in the treated and control groups of the full sample. We use the same data structure and record linkage techniques as in Chapter 2.5.1, with the only modification being the addition of x_3 to the propensity score model. We generate 100 independent simulation runs.

We define the treatment effect of interest as $\tau = \frac{1}{2}\tau_{x_3=1} + \frac{1}{2}\tau_{x_3=0} = 50$. This can be interpreted as a population average treatment effect for an x_3 that is evenly distributed across zero and one in the population, since marginally $Pr(x_3 = 1) = .5$. We estimate τ directly without estimating $\tau_{x_3=0}$ and $\tau_{x_3=1}$ separately, using (9) and (10) of the main text with each $\lambda_j = \frac{n_j}{n}$. Of course, one could estimate $\tau_{x_3=0}$ and $\tau_{x_3=1}$

as well. We leave treatment effect estimation for sub-domains using case selection procedures to future research.

The third panel of Table 7.1.4 summarizes the results. The case selection methods have smaller variance and mean squared errors than using only the known cases. The case selection methods perform similarly, with ETSR offering the smallest mean squared errors. One reason the methods perform well despite the non-constant treatment effect is that the variation in the treatment effect is not closely related to the linkage quality; in other words, the treatment effect is not systematically different for people with high versus low probabilities of being linked to their true record. We change this in Chapter 7.1.4.

Simulation with a correlation between treatment effect and link certainty

We generate the data based on the strategy in Chapter 7.1.4, but we change the generation of x_3 so that it is differentially distributed in cases that have true links and cases that do not. Specifically, for units that do not have a link across the two files, we generate x_{i3} from a Bernoulli distribution with probability 0.5. For units that have a link across the two files, we consider two subgroups. First, for the n_{c0} units where the link is uncertain (i.e., $\gamma_{fii'} = 0$ for at least one field f), we generate x_{i3} from a Bernoulli distribution with probability 0.2. Second, for the n_{c1} units where the link is certain (i.e., $\gamma_{fii'} = 1$ for all f), we generate x_{i3} from a Bernoulli distribution with probability p_c. Here, $p_c = 1 - \frac{0.5n_{c1} - 0.3n_{c0}}{n_{c1}}$. This results in a binary covariate x_3 that is more likely to equal one for units with an exact link across the files and more likely to equal zero for units without an exact link across the files, and marginally is evenly distributed across zero and one in the population. We draw each unit's treatment status w_i from a Bernoulli distribution with probability $(.4x_{i3} + .6(1 - x_{i3}))$, so that x_3 is not equally balanced in treated and control units. We generate each unit's (x_{i1}, x_{i2}) according to the distributions in Chapter 2.5.1.

Finally, we set $\tau_i = 50 + 20x_3 - 20(1 - x_3)$. We use the same analysis strategy as in Chapter 7.1.4.

The fourth panel of Table 7.7 displays the results. In this simulation, MEV performs the best among the case selection procedures. ETSR and MEDOV still offer improvements over using the known links only, although the improvements are modest compared to those for MEV. ETSR and MEDOV are conservative in choices of thresholds, which makes them have similar biases as using only the known links.

Simulation with a correlation between linking variable and outcomes

In this simulation, we generate data with a correlation between a linking variable and the outcomes with a constant treatment effect $\tau = 50$. Before we begin the linkage and causal analysis, we create a cleaned and centered version of the birth month linking variable. For any unit i, we generate $Pr(w_i = 1)$ by sampling from a Beta distribution with parameters $(birth\_month_i + 6, 12)$. As a result, the likelihood of treatment is centered at 0.5 but is lower for people with earlier birth months and higher for people with later birth months. As in Chapter 2.5.1, we generate $x_{i1} \sim Poisson(8 - 3w_i)$ and $x_2 \sim N(-w_i, 3^2)$. We generate each $y_i = 5x_{i1} + 3x_{i2} + 5birth\_month_i + \tau w_i + N(5, 10^2)$. We use Fellegi-Sunter record linkage and propensity score subclassification as before, but include birth month as a predictor in the propensity score model.

The bottom panel of Table 7.7 displays the results. In this simulation, all three criteria have smaller mean squared errors than using known links only. ETSR and MEDOV offered the largest gains.

7.2 Additional county data and address comparisons for Chapter 4

Table 7.10 shows inconsistencies in county-level records for all NC counties. "Total provisional votes" refers to the number of voters on the provisional file within each

county. "VR number listed not found" refers to cases where there is a voter registration number listed on the provisional file, but that number is not found on the voter file in that county. "Approved with no match" refers to "VR number listed not found" cases where the vote was ultimately approved.

Table 7.10: Percent inconsistencies in ID-based join by county, for all NC counties.

County Name	Total provisional votes	VR number listed not found	Approved with no match
Alamance	1105	8%	9%
Alexander	230	7%	6%
Alleghany	68	6%	6%
Anson	213	0%	2%
Ashe	93	16%	16%
Avery	146	0%	1%
Beaufort	259	4%	4%
Bertie	190	8%	8%
Bladen	174	9%	6%
Brunswick	785	8%	8%
Buncombe	1036	11%	10%
Burke	368	64%	6%
Cabarrus	1883	12%	10%
Caldwell	323	23%	23%
Camden	60	73%	17%
Carteret	394	62%	38%
Caswell	79	0%	15%
Catawba	614	40%	25%
Chatham	285	16%	15%
Cherokee	50	20%	18%
Chowan	94	7%	9%
Clay	72	11%	11%
Cleveland	718	3%	3%
Columbus	283	9%	9%
Craven	645	7%	6%
Cumberland	2733	67%	7%
Currituck	167	13%	13%
Dare	265	18%	18%
Davidson	1063	13%	12%
Davie	212	23%	23%

County Name	Total provisional votes	VR number listed not found	Approved with no match
Duplin	364	5%	5%
Durham	1926	7%	6%
Edgecombe	209	9%	9%
Forsyth	1881	11%	10%
Franklin	311	76%	17%
Gaston	1291	8%	8%
Gates	81	14%	14%
Graham	37	41%	5%
Granville	302	11%	10%
Greene	138	0%	13%
Guilford	1766	19%	18%
Halifax	335	20%	19%
Harnett	1360	5%	5%
Haywood	244	86%	27%
Henderson	243	16%	16%
Hertford	281	1%	1%
Hoke	613	7%	7%
Hyde	44	16%	16%
Iredell	726	23%	22%
Jackson	367	0%	7%
Johnston	1242	9%	9%
Jones	111	9%	9%
Lee	313	13%	13%
Lenoir	370	0%	13%
Lincoln	532	13%	13%
Macon	172	7%	7%
Madison	163	5%	5%
Martin	87	9%	9%
Mcdowell	183	0%	5%
Mecklenburg	3778	7%	6%
Mitchell	114	11%	11%
Montgomery	194	4%	4%
Moore	583	19%	19%
Nash	707	17%	13%
New Hanover	2243	10%	9%
Northampton	73	15%	15%
Onslow	1153	6%	6%
Orange	428	17%	17%
Pamlico	75	8%	8%
Pasquotank	397	5%	5%

County Name	Total provisional votes	VR number listed not found	Approved with no match
Pender	549	19%	19%
Perquimans	57	58%	11%
Person	204	13%	13%
Pitt	2256	4%	3%
Polk	65	6%	17%
Randolph	431	16%	16%
Richmond	544	13%	13%
Robeson	2067	3%	3%
Rockingham	710	9%	9%
Rowan	789	69%	20%
Rutherford	447	20%	20%
Sampson	317	6%	6%
Scotland	248	24%	15%
Stanly	506	11%	11%
Stokes	214	12%	12%
Surry	412	0%	7%
Swain	36	94%	47%
Transylvania	216	2%	2%
Tyrrell	40	15%	15%
Union	1118	13%	12%
Vance	221	3%	2%
Wake	6793	9%	8%
Warren	140	5%	4%
Washington	48	0%	17%
Watauga	542	8%	8%
Wayne	764	7%	5%
Wilkes	362	7%	7%
Wilson	316	20%	17%
Yadkin	158	0%	14%
Yancey	29	14%	14%

Figure 7.4 shows Jaro-Winkler scores for address comparisons in the Durham County linkage comparison space. The distributions are similar to the distributions of Levenshtein similarities, with lower similarities for voters between the ages of 26 and 40, and higher similarities for voters over the age of 65.

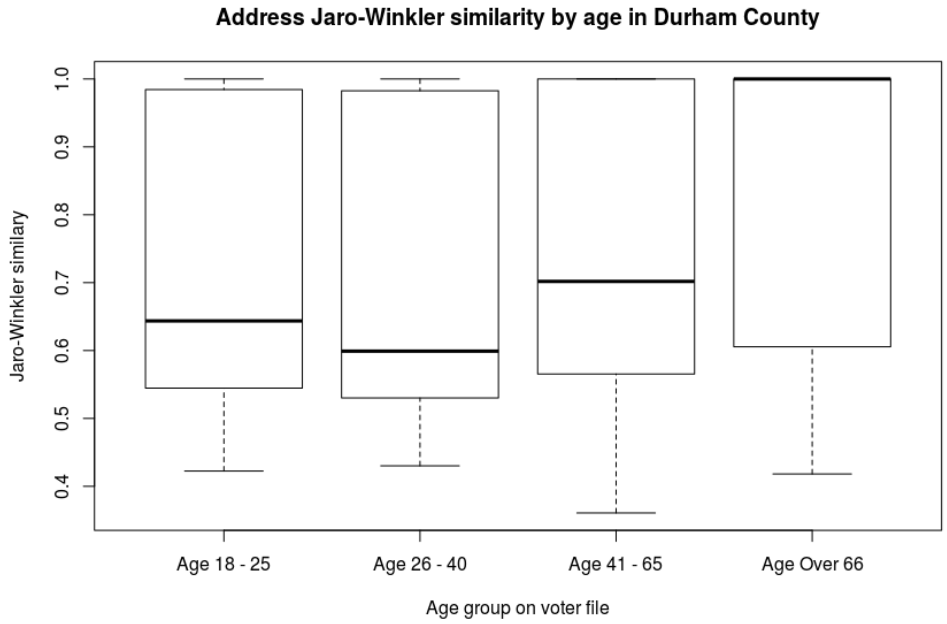

FIGURE 7.4: Durham County address Jaro-Winkler scores by age: voter registration number matches only

7.3 Additional simulation results for Chapter 5

In all of the following simulations, the match identification rate was above 97% and similar for the two approaches. We show the non-match identification rate as it more effectively shows the contrasts between accounting and not accounting for linking variable correlation.

Figure 7.5 shows the results of a simulation with the same data as in Chapter 5.4, but using the weaker prior for the combined linking variable, as shown in Table 5.4. Despite the high correlation, the decrease in the number of parameters leads to a lower non-match identification rate.

Figure 7.6 shows the non-match identification rates when the stronger prior from Table 5.4 is used, but the comparison of party affiliation given race field comparison is generated as follows.

Party affiliation given Γ_3: $\Gamma_4|\Delta_{ij} = 1, \Gamma_3 = 0 \sim \text{Multinom}(.5, .3, .2)$, $\Gamma_4|\Delta_{ij} = 1, \Gamma_3 = 1 \sim \text{Multinom}(.2, .6, .2)$, $\Gamma_4|\Delta_{ij} = 1, \Gamma_3 = 2 \sim \text{Multinom}(.1, .1, .8)$, $\Gamma_4|\Delta_{ij} =$

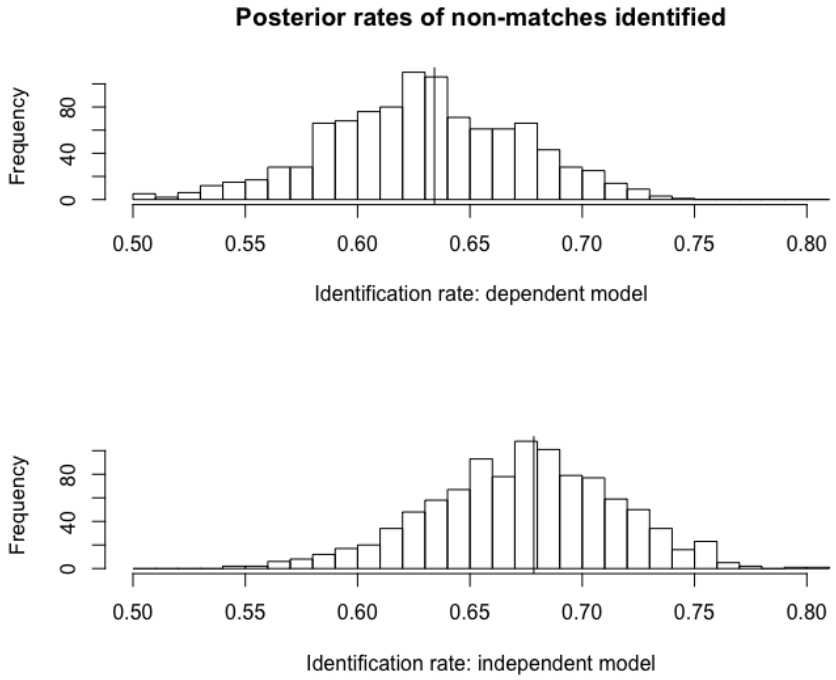

FIGURE 7.5: Posterior rate of non-match identification for the simulation using a weaker prior, accounting for and not accounting for field comparison dependence.

$0 \sim \text{Multinom}(.7, .2, .1)$

The results suggest that using a stronger prior distribution, even with a weaker dependence between race and party field comparisons, can improve classification rates.

Figure 7.7 shows the results of a simulation using the stronger prior shown in Table 5.4 but no underlying correlation between the race and party field comparisons. For this simulation, we draw the party affiliation comparison as follows:

Party: $\Gamma_4|\Delta_{ij} = 1 \sim \text{Multinom}(.1, .1, .8)$, $\Gamma_4|\Delta_{ij} = 0 \sim \text{Multinom}(.7, .2, .1)$

The results suggest that using the dependent field model with non-correlated field comparisons may not change the linkage quality if an appropriate prior distribution is specified.

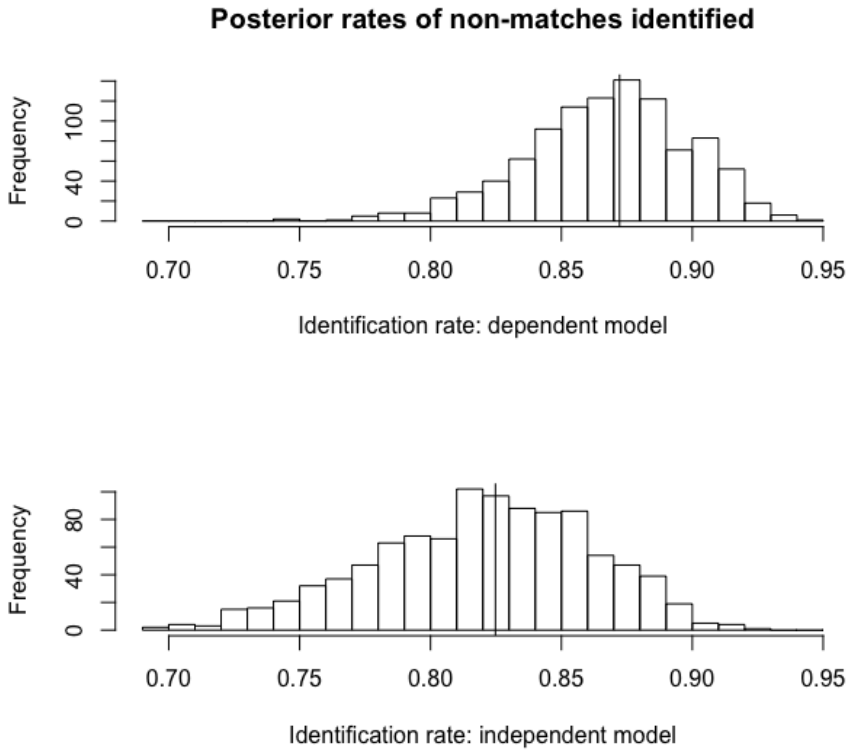

FIGURE 7.6: Posterior rate of non-match identification for the simulation with a stronger prior and weaker field comparison dependence, accounting for and not accounting for that field comparison dependence.

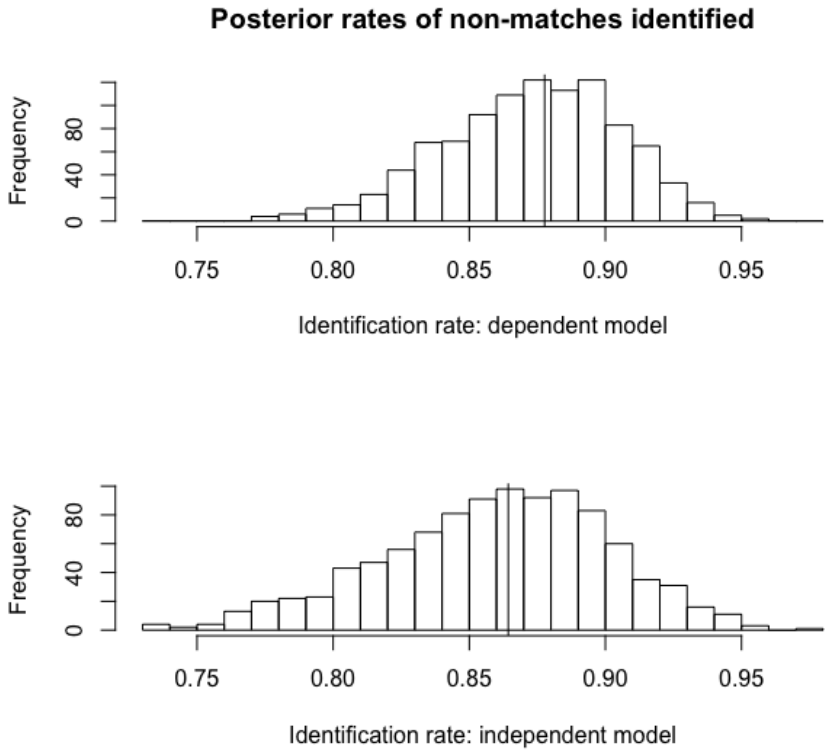

FIGURE 7.7: Posterior rate of non-match identification for the simulation with a stronger prior and underlying conditional independence of field comparisons, where model does and does not allow field comparison dependence.

Bibliography

Belin, T. and Rubin, D. (1995), "A method for calibrating false-match rates in record linkage," *Journal of the American Statistical Association*, 90, 694–707.

Borg, A. and Sariyar, M. (2015), "RecordLinkage: record linkage in R," R package version 0.4-7.

Chambers, R. (2008), "Regression analysis of probability-linked data," *Statisphere*, 4.

Chipperfield, J., Bishop, G., and Campbell, P. (2011), "Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data," *Survey Methodology*, 37, 13–24.

Christen, P. (2012), *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer Science & Business Media.

CNN Political Ticker (January 5, 2009), "Minnesota canvassing board certifies Franken win," .

D'Agostino, R. (1998), "Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group," *Statistics in Medicine*, 17, 2265–2281.

Dalzell, N. and Reiter, J. (2018), "Regression modeling and file matching using possibly erroneous matching variables," *Journal of Computational and Graphical Statistics*.

Fellegi, I. and Sunter, A. (1969), "A theory for record linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Gu, L., Baxter, R., Vickers, D., and Rainsford, C. (2003), "Record linkage: current practice and future directions," *CSIRO Mathematical and Information Sciences Technical Report*, 3, 83.

Gutman, R., Afendulis, C., and Zaslavsky, A. M. (2013), "A Bayesian procedure for file linking to analyze end-of-life medical costs," *Journal of the American Statistical Association*, 108, 34–47.

Herzog, T., Scheuren, F., and Winkler, W. (2007), *Data Quality and Record Linkage Techniques*, Springer Science & Business Media.

Imbens, G. and Rubin, D. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Jaro, M. (1989), "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Jaro, M. (1995), "Probabilistic linkage of large public health data files," *Statistics in Medicine*, 14, 491–498.

Jarvis, C. (November 10, 2016), "McCrory, Cooper gird for legal fight over votes," *The News and Observer*.

Jin, L., Li, C., and Mehrotra, S. (2003), "Efficient record linkage in large data sets," in *Proceedings of the Eighth International Conference on Database Systems for Advanced Applications*, pp. 137–146, IEEE Computer Society.

Kim, G. and Chambers, R. (2009), "Regression analysis under incomplete linkage," Tech. rep., Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 17-09.

Kim, G. and Chambers, R. (2011), "Regression analysis under probabilistic multi-linkage," Tech. rep., Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 10-11.

Lahiri, P. and Larsen, M. (2005), "Regression analysis with linked data," *Journal of the American Statistical Association*, 100, 222–230.

Larsen, M. and Rubin, D. (2001), "Iterative automated record linkage using mixture models," *Journal of the American Statistical Association*, 96, 32–41.

Levenshtein, V. I. (1966), "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet Physics Doklady*, vol. 10, pp. 707–710.

Mateyka, P. J. (2015), "Desire to move and residential mobility: 2010–2011," .

Newcombe, H., Kennedy, J., Axford, S., and James, A. (1959), "Automatic linkage of vital records," *Science*, 130, 954–959.

North Carolina State Board of Elections (2017), "About Us: North Carolina State Board of Elections," .

Rosenbaum, P. (1995), *Observational Studies*, New York: Springer.

Rosenbaum, P. and Rubin, D. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosenbaum, P. and Rubin, D. (1984), "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, 79, 516–524.

Rubin, D. (1978), "Bayesian inference for causal effects: The role of randomization," *The Annals of Statistics*, pp. 34–58.

Rubin, D. (1980), "Randomization analysis of experimental data: The Fisher randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

Rubin, D. (1990), "Comment: Neyman (1923) and causal inference in experiments and observational studies," *Statistical Science*, 5, 472–480.

Sadinle, M. (2017), "Bayesian estimation of bipartite matchings for record linkage," *Journal of the American Statistical Association*, 112, 600–612.

Sadinle, M. and Fienberg, S. E. (2013), "A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems," *Journal of the American Statistical Association*, 108, 385–397.

Scheuren, F. and Winkler, W. (1997), "Regression analysis of data files that are computer matched - part II," *Survey Methodology*, 23, 157 – 165.

Stuart, E. (2010), "Matching methods for causal inference: a review and a look forward," *Statistical Science*, 25, 1–21.

Tancredi, A. and Liseo, B. (2011), "A hierarchical Bayesian approach to record linkage and population size problems," *The Annals of Applied Statistics*, 5, 1553–1585.

Williams, P. (November 4, 2016), "Judge Says North Carolina Illegally Purged Voter Lists," *NBC News*.

Williamson, E., Morley, R., Lucas, A., and Carpenter, J. (2012), "Variance estimation for stratified propensity score estimators," *Statistics in Medicine*, 15, 1617 – 1632.

Winkler, W. (1990), "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage." .

Wortman, J. H. and Reiter, J. P. (2018), "Simultaneous record linkage and causal inference with propensity score subclassification," *Statistics in Medicine*, 37, 3533 – 3546.

Biography

Joan Pearson Heck Wortman (Jody) received her bachelors degree *magna cum laude* from Harvard University in 2014 with a concentration in Statistics and plans to receive her PhD in Statistical Science from Duke University in May 2019.

In 2016, Jody took a semester off from her PhD work to run the Data and Analytics team for Hillary Clinton's coordinated campaign in North Carolina and then was a data and analytics advisor for Roy Cooper's gubernatorial campaign later that year. She is currently the Data Science Lead at the Democratic National Committee and will continue in that role after graduation.