Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Estimating Racial Disparities When Race is Not Observed

Cory McCartan, Robin Fisher, Jacob Goldin, Daniel E. Ho & Kosuke Imai

To cite this article: Cory McCartan, Robin Fisher, Jacob Goldin, Daniel E. Ho & Kosuke Imai (2025) Estimating Racial Disparities When Race is Not Observed, Journal of the American Statistical Association, 120:552, 2140-2153, DOI: 10.1080/01621459.2025.2526695

To link to this article: https://doi.org/10.1080/01621459.2025.2526695

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 552, 2140–2153: Applications and Case Studies
https://doi.org/10.1080/01621459.2025.2526695

Estimating Racial Disparities When Race is Not Observed

Cory McCartan[a], Robin Fisher[b], Jacob Goldin[c,d], Daniel E. Ho[e], and Kosuke Imai[f]

[a]Department of Statistics, Pennsylvania State University, State College, PA; [b]Office of Tax Analysis, U.S. Department of the Treasury, Washington, DC; [c]Law School, University of Chicago, Chicago, IL; [d]NBER, Cambridge, MA; [e]Departments of Political Science and Computer Science, Law School, Stanford University, Stanford, CA; [f]Department of Government and Department of Statistics, Harvard University, Cambridge, MA

ABSTRACT

Estimating racial disparities without access to individual-level racial information is a common challenge in economic and policy settings. We develop a statistical method that relaxes the strong independence assumption of common race imputation approaches like Bayesian-Improved Surname Geocoding (BISG). Our identification assumption is that surname is conditionally independent of the outcome given (unobserved) race, residence location, and other observed characteristics. The proposed approach reduces error by up to 84% relative to BISG when estimating racial differences in political party registration. In our application, we estimate racial differences in who benefits from the home mortgage interest deduction using individual-level tax data from the U.S. Internal Revenue Service. Our analysis reveals that many fewer Black and Hispanic filers claim the HMID than White and Asian filers. We also find that the racial gaps in homeownership rates alone cannot explain this disparity. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received December 2024
Accepted June 2025

KEYWORDS
BISG; Ecological inference; Instrumental variables; Proxy variable; Race imputation

1. Introduction

The identification and estimation of racial disparities is of paramount importance to researchers, policymakers and organizations in a variety of areas including public health, employment, voting, criminal justice, economic policy and taxation, housing, lending, and technology and fairness. In many cases, however, racial information is not available at the individual level. The unavailability of individual racial information makes it impossible for analysts to simply tabulate variables of interest against race to identify disparities among different racial groups. In fact, in some areas, the law explicitly prohibits the collection of racial information even as it demands fair treatment on the basis of race (see, e.g., the U.S. Equal Credit Opportunity Act). This creates a dilemma for organizations who wish to measure possible disparities in order to monitor the fairness of their decision-making or service provision.

In our empirical application, we analyze large-scale administrative tax data from the U.S. Internal Revenue Service (IRS), which does not collect individual taxpayers' racial information. As briefly described in Section 2, our goal is to estimate the distribution by race in who claims the home mortgage interest deduction (HMID). The HMID is one of the largest tax benefits for homeowners in the income tax code, and some scholars have claimed it disproportionately benefits taxpayers in certain racial groups (Moran and Whitford 1996; Brown 2022). We investigate this question, whose answer has been largely hampered by a lack of administrative tax data with taxpayers' racial information.

To estimate racial disparities without individual racial data, some researchers have turned to ecological inference methods (e.g., Goodman 1953; King 1997; Wakefield 2004; Imai, Lu, and Strauss 2008). These methods, however, require strong assumptions, which are difficult to verify and may provide misleading results (Cho and Manski 2008). They also rely on accurate marginal information about race, which may not be available.

Where the analysis of racial disparities involves large-scale administrative data, many analysts have adopted Bayesian Improved Surname Geocoding (BISG), which generates individual probabilities of belonging to different racial groups using Bayes' rule applied to last names and geographic location (Fiscella and Fremont 2006; Elliott et al. 2008; Imai and Khanna 2016). BISG leverages residential racial segregation and the association between self-reported race and surname to produce generally accurate and calibrated predictions of self-reported individual race (Kenny et al. 2021; DeLuca and Curiel 2022).

Much attention has been given to ways of increasing the accuracy of BISG and related methods for race *prediction* (Voicu 2018; Zest AI 2020; Argyle and Barber 2024; Imai, Olivella, and Rosenman 2022; Decter-Frain 2022; Greengard and Gelman 2023). Unfortunately, accurate BISG racial prediction alone does not guarantee the unbiased estimation of racial *disparities*, which is the ultimate goal of most analysts. To estimate disparities, BISG probabilities (or any other racial predictions) must be combined with information on the outcome variable for which the disparities are of interest. But the most common techniques are known to be biased when race is correlated with the outcome even after controlling on name and location (Chen et al.

CONTACT Kosuke Imai *imai@harvard.edu* Department of Government and Department of Statistics, Harvard University, Cambridge, MA 02138.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

2019; Argyle and Barber 2024; Greenwald et al. 2023). These approaches include *weighting* the outcome variable by the BISG probabilities, and *thresholding* the BISG probabilities to produce point predictions of individual race.

In fact, these methods often *underestimate* racial disparities, which is problematic for policymakers and analysts who aim to identify these disparities. As discussed in Section 3, the standard methods of racial disparity estimation based on BISG predictions implicitly require individuals' race to be conditionally independent of the outcome given their residence location, surnames, and other observable attributes. This key assumption, however, is violated if race affects many aspects of society even after accounting for residence location, surnames, and other observable attributes. Researchers have noted the implausibility of this assumption and have advocated for partial identification strategies (Kallus, Mao, and Zhou 2022; Elzayn et al. 2023). Another literature considers general problems of mismeasurement and develops methodology for the case where a randomly-sampled validation set containing true labels is available (Selén 1986; Fong and Tyler 2021; Egami et al. 2024; Angelopoulos, Duchi, and Zrnic 2024). However, in many settings, such as our motivating application, ground-truth racial information is not available for even a subset of the records.

To address this challenge, in Section 4, we propose an alternative identification strategy. Specifically, we assume that the outcome is conditionally independent of surname given (unobserved) individual's race, residence location, and other observed attributes. This assumption is a type of exclusion restriction where surname serves as an instrumental variable for unobserved race. It implies that for two individuals who live in the same area, belong to the same racial group, and share the observable attributes, their surnames have no predictive power of the outcome. Somewhat counter-intuitively, the high-dimensionality of surnames aids rather than hinders identification because it provides a large number of instruments. We argue that this new identification assumption is more credible than the commonly invoked assumption unless surname is directly used to determine the outcome of interest (i.e., name-based discrimination).

Leveraging this identification strategy, in Section 4.2 we introduce a new class of models, Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), that accurately estimates racial disparities using BISG probabilities. Beyond accuracy, BIRDiE improves on standard methodology in several ways:

- BIRDiE includes built-in flexibility for researchers to make problem-specific modeling choices (Section 4.2).
- BIRDiE can be fit with an EM algorithm that can scale to hundreds of thousands or millions of observations (Appendix B.2).
- BIRDiE produces updated BISG probabilities that incorporate the outcome variable and are likely to be more accurate than the BISG probabilities based only on surnames and geolocation (Section 4.3).
- BIRDiE can condition on additional variables whose distribution by race is not known a priori (Section 4.4). For example, party identification can be estimated by race *and* turnout.

Finally, in Section 4.5 we address potential violations of the key identification assumption, such as the one caused by overly coarse racial categories, by exploiting auxiliary information about the relations between names and more specific ethnic groups. All of the proposed methodology is implemented in a computationally efficient open-source software package, `birdie`, which is available on CRAN at *https://CRAN.R-project.org/package=birdie*.

In Section 5, we validate the proposed methodology using the voter file in North Carolina, where self-reported individual race is used to construct the ground-truth of racial disparities. BIRDiE substantially outperforms existing estimators across multiple levels of geolocation specificity. For example, the most popular existing BISG-only disparity estimator pegs the gap at Democratic party registration between White and Black voters at 16.8 percentage points (pp), while the actual gap is 54.6pp—more than double. Our preferred BIRDiE model using the same BISG probabilities yields an estimate of 48.8pp. This represents about a 85% reduction in bias.

In Section 6, we apply BIRDiE to large-scale administrative tax data from the U.S. Internal Revenue Service, which does not collect individual taxpayers' racial information. We produce novel estimates of the distribution by race in who claims the home mortgage interest deduction—a question that has largely been hampered by a lack of administrative tax data with taxpayers' racial information. Our analysis reveals a substantial degree of racial disparity with many fewer Black and Hispanic filers claiming the HMID than White and Asian filers. We find that the racial gaps in homeownership rates alone cannot explain this disparity. Section 7 gives concluding remarks.

2. Racial Disparity in Home Mortgage Interest Deduction

In this section, we briefly describe our empirical application—estimation of racial disparity in HMID. The HMID is designed to incentivize home ownership; homeowners with a mortgage qualify for an itemized deduction based on the amount of mortgage interest they pay during the year. The deduction is only available for taxpayers who itemize their deductions. Following increases to the standard deduction and other tax code changes as part of the "Tax Cuts and Jobs Act" of 2017 (P.L. 115-97), roughly 90% of taxpayers take the standard deduction and do not itemize. These taxpayers are unable to take advantage of the HMID.

The Treasury Department estimates the HMID costs the government about $25 billion in foregone revenue in 2019. By budgetary cost, the deduction is the largest in the income tax code (Congressional Research Service 2017). Because it is only available to homeowners, the HMID may disproportionately benefit taxpayers of racial groups that have a high homeownership rate. Prominent legal scholars have criticized the possible disproportionate benefits, with Brown (2022, p. 94) referring to the subsidy from the HMID as "little more than the twenty-first-century version of redlining" and concluding it "must be repealed." On the other hand, it is also possible that if Black homeowners faced higher mortgage rates, they could in principle benefit more than would be expected based on homeowner-

ship rates alone. Lack of administrative data on HMID claims by race made it difficult to quantify racial disparities that potentially exist.

The Treasury's internal Office of Tax Analysis has recently used an extension of the standard BISG model to estimate the usage of the HMID and other deductions by race from individual-level data (Cronin, DeFilippes, and Fisher 2023). External researchers have also studied HMID usage by analyzing survey data or data on proxies like home ownership (Sullivan et al. 2017). Both types of analyses have found that White taxpayers benefit disproportionately from the HMID, though the magnitude of the disparity is unclear, especially given the methodological challenges identified in this article.

In Section 6, we apply BIRDiE to more precisely answer the question of which groups are using the HMID and how much they benefit from it. We analyze individual-level tax data from the IRS, which includes the universe of income tax returns filed by U.S. taxpayers. The IRS does not collect information on taxpayer race or ethnicity. While, unlike in many settings, it may theoretically be possible to link tax data to census data that contain race, such linkages are often prohibited by law, such as Titles 13 and 26 of the U.S. Code. Researchers are not currently permitted to link these specific Treasury data to Census Bureau data with individual racial identifiers.

3. Bias of the Standard Methodology

In this section, we review the assumptions of the standard BISG-based methodology for estimating racial disparities when individual race is not observed. We show that these racial disparity estimates are biased unless the outcome variable is independent of race given surname, residence location, and other observed covariates. We argue that this assumption is likely to be violated given the significant role race plays in our society.

3.1. Setup and BISG Procedure

Suppose that we have an iid sample of N individuals from a super population. For each individual $i = 1, \ldots, N$, we define a tuple $(Y_i, R_i, G_i, X_i, S_i)$, where $Y_i \in \mathcal{Y}$ is the outcome, $R_i \in \mathcal{R}$ is the (unobserved) race of the individual, $G_i \in \mathcal{G}$ is the (geo)location of the individual's residence, $X_i \in \mathcal{X}$ are other observed characteristics, and $S_i \in \mathcal{S}$ is the individual's surname. When we are not referring to a particular individual, we will drop the subscripts for simplicity. Note that individual race is unobservable but all other variables are assumed to be observed. The availability of particular (or any) X is not required for either the standard or proposed methodology.

We assume throughout that these variables are discrete, taking a finite set of values, that is, $|\mathcal{Y}|$, $|\mathcal{R}|$, $|\mathcal{G}|$, $|\mathcal{X}|$, and $|\mathcal{S}|$ are constants. Note that typically S is high-dimensional as there exist a large number of unique surnames. In practice, residence location G is also discrete, since joint information about location, race, and other variables is generally only available down to the Census block level. For simplicity, we assume that the outcome variable Y is also discrete, though it is possible to extend the standard and proposed methodologies to continuous outcome variables.

Typically, BISG relies on data from the decennial Census or the American Community Survey (ACS), which provide information on the joint distribution of R and G (and any other covariates X, such as gender or age). It then combines this information with data from the Census Bureau's surname tables (U.S. Census Bureau 2014), which provide information on the joint distribution of R and S. We summarize this set of information from the Census by two conditional probabilities, $\mathbf{q}_{GX|R}$ and $\mathbf{q}_{S|R}$, and one marginal probability, $\mathbf{q}_R$.

The BISG estimator of the probability that individual i belongs to race $r \in \mathcal{R}$ can then be written as (Fiscella and Fremont 2006; Elliott et al. 2008)

$$\widehat{P}_{ir} \coloneqq \frac{q_{G_iX_i|r}\, q_{S_i|r}\, q_r}{\sum_{r' \in \mathcal{R}} q_{G_iX_i|r'}\, q_{S_i|r'}\, q_{r'}}, \tag{1}$$

where, for example, $q_{G_iX_i|r}$ indicates the estimated conditional probability of residence location G_i and covariates X_i given race r, taken from the Census table $\mathbf{q}_{GX|R}$.

The BISG estimator relies on two key assumptions. The first is that the Census tables reflect the true population distributions of R, S, G, and X.

Assumption ACC (Data accuracy). For all i, we have:

$$\begin{aligned} \mathbb{P}(R_i = r) &= q_r, \\ \mathbb{P}(S_i = s \mid R_i = r) &= q_{s|r}, \\ \mathbb{P}(G_i = g, X_i = x \mid R_i = r) &= q_{gx|r}. \end{aligned}$$

Despite the best efforts of the Census Bureau, Assumption ACC may never hold exactly in practice. The decennial census has intrinsic error, including undercounting minority groups (U.S. Census Bureau 2022; Anderson and Fienberg 1999; Strmic-Pawl, Jackson, and Garner 2018) and error introduced by privacy-preserving mechanisms (Kenny et al. 2024). And because of births, deaths, and moves, census data are often out-of-date from the moment of publication. These errors have led further extensions of the BISG estimator to account for measurement error (Imai, Olivella, and Rosenman 2022).

The plausibility of Assumption ACC is stretched further when the study population is a subset of the whole U.S. population, and so is not covered by national census data. In these cases, analysts should set $\mathbf{q}_R$ to the known or estimated marginal racial distribution in the study population, rather than the national racial distribution. It may be more plausible then to assume that the conditional distributions $\mathbb{P}(S \mid R)$ and $\mathbb{P}(G, X \mid R)$ match the census distributions, even if $\mathbb{P}(R)$ does not (Rosenman, Olivella, and Imai 2023). Greengard and Gelman (2023) take this approach a step further by raking BISG probabilities to all known margins, further improving calibration.

The second assumption required by BISG is the following conditional independence relation between an individual's surname and residence location (as well as other characteristics) given their unobserved race.

Assumption CI-SG (Conditional independence of name and other proxy variables). For all i,

$$S_i \perp\!\!\!\perp \{G_i, X_i\} \mid R_i.$$

Assumption CI-SG implies, for example, that once we know an individual is White, knowing their surname is Smith tells us nothing about their residence location and other observed characteristics. Although this assumption appears to be reasonable, the lack of granularity in the coding of race may lead to its violation. For example, people with Chinese, Indian, Filipino, Vietnamese, Korean, or Japanese are all coded as one racial group "Asian" in the census. These groups, however, have various surnames and have different demographic and geographic distributions. For instance, unlike the Smith example, knowing that an Asian individual's surname is Gupta makes it more likely that they have a higher income and live in the Eastern U.S (Budiman, Cilluffo, and Ruiz 2019).

Even though Assumptions CI-SG and ACC may not hold exactly, researchers find that BISG produces accurate and generally well-calibrated estimates in practice (Imai and Khanna 2016; Zhang 2018; Kenny et al. 2021; DeLuca and Curiel 2022). We observe this pattern as well in the validation study in Section 5. Under Assumption CI-SG, by Bayes' Rule,

$$\begin{aligned}&\mathbb{P}(R_i = r \mid G_i, X_i, S_i)\\ &\quad \propto \mathbb{P}(G_i, X_i \mid R_i = r)\mathbb{P}(S_i \mid R_i = r)\mathbb{P}(R_i = r).\end{aligned}$$

This justifies the estimator given in (1), yielding the following immediate result.

Proposition 3.1 (Accuracy of BISG). Under Assumptions CI-SG and ACC, the BISG estimator produces correct probabilities. That is, we have $\widehat{P}_{ir} = \mathbb{P}(R_i = r \mid G_i, X_i, S_i)$.

New methods are developed to improve the calibration of BISG probabilities, including some machine learning methods based on labeled data (Zest AI 2020; Imai, Olivella, and Rosenman 2022; Argyle and Barber 2024; Decter-Frain 2022; Greengard and Gelman 2023; Cheng et al. 2023). Fundamentally, these approaches all focus on building a more accurate model for $R \mid G, X, S$ at the individual level.

3.2. Bias of BISG-based Racial Disparity Estimates

To estimate racial disparities, BISG probabilities (or other racial predictions) must be combined with the outcome variable. There are several common ways researchers do this. The most frequent is the *thresholding* or *classification* estimator, which deterministically assigns individuals to a predicted racial category based on the BISG estimates $\widehat{\mathbf{P}}_i$ (either the largest $\widehat{P}_{ir}$ or the one which exceeds a predetermined threshold). Estimates of $\mathbb{P}(Y = y \mid R = r)$ are then obtained by tabulating the data by these assigned categories. Another common approach, which attempts to capture the uncertainty inherent in race prediction, is the following *weighting* estimator:

$$\widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid r) = \frac{\sum_{i=1}^{N} \mathbf{1}\{Y_i = y\}\widehat{P}_{ir}}{\sum_{i=1}^{N} \widehat{P}_{ir}}.$$

Unfortunately, accurate and calibrated predictions of individual race alone are not sufficient for unbiased estimation of racial disparities using these standard methodologies. This should come as no surprise for the thresholding estimator, since it does not take into account prediction uncertainty in the BISG probabilities. This is akin to ignoring measurement errors in the covariates of a regression, something which has long been known to lead to biased coefficient estimates. Unlike the classical errors-in-variables setting, however, the bias of the thresholding estimator is not consistently in the same direction, making it hard to reason about (Chen et al. 2019).

But, the weighting estimator is also biased because the prediction error of race probabilities may be correlated with the outcome variable of interest. Fortunately, unlike the threshold estimator, it is easier to understand the nature of this bias. Chen et al. (2019) show that the asymptotic bias of the weighting estimator is controlled by the residual correlation of Y and R after adjusting for G, X, and S. We reproduce this result here.

Theorem 3.2 (Theorem 3.1 of Chen et al. 2019). If race is binary ($\mathcal{R} = \{0, 1\}$), then as $N \to \infty$,

$$\begin{aligned}&\widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid r) - \mathbb{P}(Y = y \mid R = r)\\ &\quad \xrightarrow{a.s.} -\frac{\mathbb{E}[\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = r\} \mid G, X, S)]}{\mathbb{P}(R = r)}.\end{aligned}$$

This result implies that when the BISG residuals $\mathbf{1}\{R = r\} - \mathbb{P}(R = r \mid G, X, S)$ are correlated with the outcome, estimates will be biased. In fact, the weighting estimator will often underestimate the magnitude of a disparity, as the following corollary shows. For instance, in measuring disparities in loan approval, if Black applicants are less likely to be approved for loans across all locations and surnames than White applicants, then the weighting estimator would understate the resulting overall White-Black disparity in loan approval rates.

Corollary 3.2.1 (Underestimation of racial disparity). Let $y \in \mathcal{Y}$. If race is binary ($\mathcal{R} = \{0, 1\}$), and $\mathbb{P}(Y = y \mid R = 1, G = g, X = x, S = s) > \mathbb{P}(Y = y \mid R = 0, G = g, X = x, S = s)$ for all $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$, then

$$\begin{aligned}&\widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid 1) - \widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid 0)\\ &\quad < \mathbb{P}(Y = y \mid R = 1) - \mathbb{P}(Y = y \mid R = 0).\end{aligned}$$

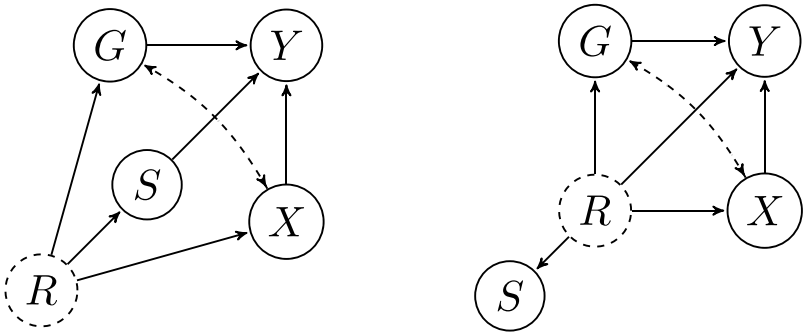
Conversely, as formally stated below, Theorem 3.2 implies that conditional independence between an individual's race and outcome given their surname, residence location, and other characteristics is sufficient to eliminate the asymptotic bias of the weighting estimator.

Assumption CI-YR (Conditional independence of outcome and race). For all i,

$$Y_i \perp\!\!\!\perp R_i \mid G_i, X_i, S_i.$$

Corollary 3.2.2 (Consistency of weighting under Assumption CI-YR). Let $y \in \mathcal{Y}$. If race is binary (so $\mathcal{R} = \{0, 1\}$), and Assumption CI-YR holds, then as $N \to \infty$, $\widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid r) - \mathbb{P}(Y = y \mid R = r) \xrightarrow{a.s.} 0$.

Figure 1(a) shows a causal directed acyclic graph (DAG) that satisfies Assumptions CI-YR and CI-SG. The dashed node border for R represents the fact that race is unobserved. The causal structure in Figure 1(a) implies the conditional independence relation $Y \perp\!\!\!\perp R \mid G, X, S$, because all paths from R to Y are



(a) **Assumption CI-YR**: Conditional independence of Y and R given (G, S, X)

(b) **Assumption CI-YS**: Conditional independence of Y and S given (G, R, X)

Figure 1. Possible causal structures for which each of the labeled assumptions is satisfied, represented as a directed acyclic graph (DAG) where *G* is residence location, *R* is race, *S* is surname, *X* is observed covariates, and *Y* is the outcome. Race *R* is unobserved, which is signified by a dashed node boundary. Both DAGs also satisfy Assumption CI-SG: Conditional independence of *S* and (*G*, *X*) given *R*.

blocked by *G*, *X*, or *S*. Thus, this DAG assumes that the effect of race *R* on the outcome *Y* must be entirely mediated by surname *S*, residence location *G*, and other observed characteristics *X*. We note that causal structures other than Figure 1(a) may likewise imply Assumption CI-YR. Assumption CI-YR is not credible in many real-world settings because race can affect the outcome through so many factors, biasing the weighting estimator.

In other settings, Assumption CI-YR may be plausible. For example, a manager who reviews job applications without individual-race information may be influenced by racial or gender cues in an applicant's name or address (Park et al. 2009; Åslund and Skans 2012). So long as we observe all information used by the manager and incorporate it in the BISG estimation, the weighting estimator would be asymptotically unbiased. Similarly, in evaluating the fairness of algorithmic decision-making, as long as all the information used by the algorithm is incorporated into BISG, Assumption CI-YR would be appropriate. Outside of these cases, however, the weighting estimator is likely to be biased.

Finally, while our discussion has been focused on the BISG methodology, the results and necessary assumptions carry over to other approaches which produce probabilistic predictions of individual race (Zest AI 2020; Argyle and Barber 2024; Imai, Olivella, and Rosenman 2022; Decter-Frain 2022; Greengard and Gelman 2023). Just like standard BISG, all of these methods are based on individual names, geographic location (and sometimes other geographic attributes), and possibly additional individual covariates. Thus, well-calibrated probabilities are not generally sufficient to produce unbiased estimates of racial disparities using the standard weighting or thresholding estimators.

4. The Proposed Methodology

In this section, we propose an alternative identification strategy that allows race to directly affect the outcome of interest. We show that racial disparity is identifiable if surname is conditionally independent of the outcome given race, residence geolocation, and other observed information, and the aforementioned assumptions required by BISG hold.

We develop a class of statistical models, called Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), that estimate racial disparity under this identification condition by using surnames as an instrumental variable for race. BIRDiE models take as inputs the BISG probabilities, and so can be easily applied on top of existing analysis pipelines, including those with alternative probabilistic race prediction methodologies. We also discuss computational strategies to handle large datasets, and an extension of the methodology to include an additional explanatory variable that was not used at the BISG stage. Finally, we show how to address potential violations of the key identification assumptions, such as those caused by name-based discrimination.

4.1. New Identification Strategy

To reduce the potential bias of the weighting estimator, we propose an alternative identification assumption that may be applicable when Assumption CI-YR is not credible. Specifically, we assume that surname, rather than race, satisfies the exclusion restriction conditional on (unobserved) race, residence location, and other observed characteristics.

Assumption CI-YS (Conditional independence of outcome and name). For all *i*,

$$Y_i \perp\!\!\!\perp S_i \mid R_i, G_i, X_i.$$

Figure 1(b) shows one possible causal DAG that meets this assumption as well as Assumption CI-SG. In this DAG, race can have a direct effect on the outcome *Y* as well as on residence location *G* and other observed characteristics *X*, while all paths from *S* to *Y* are blocked by *G*, *X*, or *R*.

This causal structure is often more plausible than Assumption CI-YR because Assumption CI-YS allows race to directly affect the outcome. For party registration, Assumption CI-YS implies that among White voters in a particular geographic region, voters named Smith would *a priori* no more or less likely to identify with one party than voters named Thomas. In contrast, Assumption CI-YR would mean that among voters named Smith in a particular geographic region, White voters would be *a priori* be no more or less likely to identify with

one party than Black voters. In this case, Assumption CI-YR is likely to be violated, while Assumption CI-YS is plausible. We emphasize a key trade-off between the two assumptions. While Assumption CI-YS rules out the possibility that surname directly affects the outcome (e.g., name-based discrimination), such a direct effect is allowed under Assumption CI-YR. Section 4.5 revisits this important issue.

The validity of each assumption depends on a specific application. In the above hiring example, if the manager reviews applicants anonymously, there will be no name-based discrimination and the assumption is likely to be satisfied. The assumption may be violated in other contexts, however. For example, in studying turnout, if campaigns use the surnames of voters to decide whether to mobilize them, Assumption CI-YS will be violated.

Another possible violation of the assumption is the existence of an unobserved confounder that affects both outcome and surname. The country of origin for an immigrant may represent such a confounder. Since surnames are informative of country of origin, even within racial groups (as is often the case for Asian individuals), variations in outcomes by country of origin will likely violate Assumption CI-YS. This reflects the limitations of the relatively coarse racial classifications used in BISG, as discussed above. Even in these cases, however, conditioning on (unobserved) race is likely to substantially reduce the magnitude of association between outcome and surnames. In Section 4.5, we show how to address this potential violation of Assumption CI-YS.

The following theorem formally shows that it is possible to point-identify racial disparities under Assumption CI-YS. All proofs appear in the appendix.

Theorem 4.1 (Identification). For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then, under Assumption CI-YS, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, G = g, X = x)$ are identified if and only if both $\mathbf{P}$ and the augmented matrix $(\mathbf{P} \quad \mathbf{b})$ have rank $|\mathcal{R}|$.

The essence of this identification result is the following simple observation. Under Assumption CI-YS, we have, for all $y \in \mathcal{Y}$, $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\begin{aligned}\mathbb{P}(Y = y \mid G = g, X = x, S = s) &= \sum_{r\in\mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x) \\ &\quad \times \mathbb{P}(R = r \mid G = g, X = x, S = s). \qquad (2)\end{aligned}$$

The leftmost term is estimable from the data and corresponds to the vector $\mathbf{b}$ in Theorem 4.1, while the rightmost term is the BISG estimand and corresponds to the matrix $\mathbf{P}$. Lastly, the remaining term in the middle can be solved for, since (2) holds across all combinations of Y, G, X, and S, leading to a large system of linear equations with $(|\mathcal{Y}| - 1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{S}|$ equations and $(|\mathcal{Y}| - 1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{R}|$ unknowns. Since $|\mathcal{R}| \ll |\mathcal{S}|$, we can identify these unknowns as long as the linear system has sufficient rank. If surnames are only weakly predictive of race for some groups, then this rank condition may be threatened, or finite-sample variance may be inflated. We discuss a metric for identifying this case in Appendix C.4.2. Our result is related to causal effect identification based on the use of proxy variables for unmeasured confounding variables (Kuroki and Pearl 2014; Miao, Geng, and Tchetgen Tchetgen 2018; Knox, Lucas, and Cho 2022). This literature uses a similar identification strategy based on linear systems. Here, we use surname as a proxy variable for (unobserved) race, which is analogous to the unmeasured treatment variable in causal inference.

Together with Proposition 3.1, Theorem 4.1 implies that racial disparities can be identified under Assumptions CI-SG, ACC, and CI-YS . The identifying equation (2) shows that $\mathbb{P}(Y = y \mid G = g, X = x, S = s)$ is linear in the BISG estimands $\mathbb{P}(R = r \mid G = g, X = x, S = s)$. Thus, it is natural to consider the following least-squares estimator of $\mathbb{P}(Y = y \mid R, G = g, X = x)$ under this alternative identification strategy,

$$\widehat{\mu}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x) = (\widehat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)}\widehat{\mathbf{P}}_{\mathcal{I}(xg)})^{-1}\widehat{\mathbf{P}}_{\mathcal{I}(xg)}\,\mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\},$$

where as above $\widehat{\mathbf{P}}$ is the matrix of BISG probabilities, and $\mathcal{I}(xg)$ is the set of individuals i with $X_i = x$ and $G_i = g$. Here and throughout the article, a dot will indicate a vector constructed over that index, so $\widehat{\mu}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x)$ is a vector of conditional probabilities for a particular outcome level y across all racial groups in $\mathcal{R}$. This estimator is closely related to the two-sample instrumental variables approach of Angrist and Krueger (1992), though the required assumptions are slightly different (see also Crossley, Levell, and Poupakis (2022)).

Post-stratifying this estimator across the (G, X) cells yields an estimator of $\mathbb{P}(Y = y \mid R)$,

$$\widehat{\mu}^{(\text{p-ols})}_{Y|R}(y \mid r) = \sum_{x\in\mathcal{X}, g\in\mathcal{G}} (\widehat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)}\widehat{\mathbf{P}}_{\mathcal{I}(xg)})^{-1}\widehat{\mathbf{P}}_{\mathcal{I}(xg)}\,\mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\})_r q_{gx|r},$$

since $q_{gx|r} = \mathbb{P}(G = g, X = x \mid R = r)$ under Assumption ACC. This estimator is unbiased, as the following theorem shows.

Theorem 4.2 (Unbiasedness of OLS Estimator). If Assumptions CI-SG, ACC, and CI-YS hold, and the identification conditions in Theorem 4.1 are satisfied, then for all $y \in \mathcal{Y}$ and $r \in \mathcal{R}$,

$$\mathbb{E}[\widehat{\mu}^{(\text{p-ols})}_{Y|R}(y \mid r)] = \mathbb{P}(Y = y \mid R = r).$$

Comparing this OLS estimator with the weighting estimator $\widehat{\mu}^{(\text{wtd})}_{y|r}$ demonstrates the relationship between Assumptions CI-YR and CI-YS. The next theorem shows that within the (G, X) cells these two estimators are guaranteed to disagree, unless either the BISG probabilities perfectly discriminate or the weighting estimator is constant across races. These two conditions are almost never met in practice. This underscores the importance of selecting the appropriate assumption (CI-YR or CI-YS) for a particular analysis, since they yield different results.

Theorem 4.3 (Necessary and Sufficient Condition for Equality of the Weighting and OLS Estimators). For any $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and

$x \in \mathcal{X}$, within the set of individuals with $G_i = g$ and $X_i = x$, we have that $\widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid \cdot) = \widehat{\mu}^{(\text{ols})}_{Y|R}(y \mid \cdot)$ if and only if for every pair $j, k \in \mathcal{R}$, either the BISG probabilities perfectly discriminate (i.e., $\mathbb{P}(R_i = j \mid G_i, X_i, S_i) > 0$ implies $\mathbb{P}(R_i = k \mid G_i, X_i, S_i) = 0$ and vice versa) or $\widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid j) = \widehat{\mu}^{(\text{wtd})}_{Y|R}(y \mid k)$.

Despite potential advantages over the weighting estimator, the OLS estimator ignores the fact that the unknown parameters are probabilities and thus constrained to be nonnegative and sum to 1. As a result, in any particular sample, the estimator can produce impossible or contradictory estimates. This problem is exacerbated by the high variance that occurs when G and X partition the sample into many small cells (e.g., G represents Census tracts). Our proposed methodology in the next section incorporates this constraint and performs shrinkage across (G, X) cells, and thus outperforms the OLS estimator (see Section 5). Nevertheless, the OLS estimator is simple to implement and can perform well when each (G, X) cell has plenty of data. It also allows for the use of a variance inflation metric, which can flag cases when surnames are particularly uninformative, and thus $\widehat{\mu}^{(\text{p-ols})}_{Y|R}$ will have high variance. The use of this metric is illustrated in Appendix C.4.2. alongside a study of the finite-sample behavior of the OLS estimator.

4.2. Bayesian Instrumental Regression for Disparity Estimation

BIRDiE combines a user-specified complete-data outcome model $\pi(Y \mid R, G, X, \Theta)$, parameterized by Θ, with the BISG model in order to estimate the distribution $Y \mid R$ that is of interest. In this regard, it mirrors the two-stage instrumental variables (IV) regression: a first stage (BISG) that estimates the relationship between instrument (surname) and variable of interest (race), and a second stage (BIRDiE) that uses the first-stage estimates to produce valid estimates of the quantity of interest. Unlike two-stage IV, however, the BIRDiE approach is based on a coherent joint distribution of data, unknown parameters, and race. We exploit this fact and develop the general BIRDiE modeling approach below.

Specifically, the BIRDiE posterior is obtained by applying Assumptions CI-SG, ACC, and CI-YS to the joint distribution $\pi(Y, R, G, X, S, \Theta)$:

$$\begin{aligned}\pi(\Theta, \mathbf{R} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) &\propto \pi(\Theta) \prod_{i=1}^{N} \pi(Y_i \mid R_i, G_i, X_i, \Theta) \\ &\quad \times \pi(R_i \mid G_i, X_i, S_i) \\ &= \pi(\Theta) \prod_{i=1}^{N} \pi(Y_i \mid R_i, G_i, X_i, \Theta) \widehat{P}_{iR_i}. \quad (3)\end{aligned}$$

As above, $\widehat{\mathbf{P}}_i$ are the BISG probability estimates for individual i that depend on Census data ($\mathbf{q}_{GX|R}$, $\mathbf{q}_{S|R}$, $\mathbf{q}_R$), but not on the outcome-model parameters Θ. Because these "first-stage" BISG estimates are exact probabilities, by Proposition 3.1, they can be plugged into the BIRDiE posterior computation without losing Bayesian coherency. In practice, this procedure remains approximately valid even if a more complex model is used in place of the BISG probabilities (e.g., Zest AI 2020; Imai, Olivella, and Rosenman 2022; Argyle and Barber 2024; Decter-Frain 2022).[1]

Application of this general BIRDiE approach requires choosing a complete-data outcome model, given by the likelihood $\pi(Y_i \mid R_i, G_i, X_i, \Theta)$ and prior $\pi(\Theta)$. Since Y is discrete, a categorical regression model is appropriate for $\pi(Y_i \mid R_i, G_i, X_i, \Theta)$. The exact model specification will depend on the analyst's goals, computational resources, and prior beliefs about the structure of the problem. Appendix B.1 presents several reasonable alternatives that trade off modeling flexibility and computational efficiency.

There is also a computational challenge in applying BIRDiE, because the posterior in (3) contains the high-dimensional discrete nuisance parameter $\mathbf{R}$. This challenge is compounded for larger sample sizes. Our primary recommendation is an EM algorithm for fitting the model, which we derive and explain in Appendix B.2. Other computational approaches discussed in the appendix include marginalizing out $\mathbf{R}$ and using a Gibbs sampler, which is closely related to the EM algorithm. A critical advantage of the proposed EM scheme over working with the marginal likelihood or directly with the full posterior is that the maximization in the M-step can be performed using sufficient statistics calculated as part of the E-step, rather than on all of the individual entries in the data. Since the M-step is usually the bottleneck in the computation, this is enormously helpful—the problem size scales with $|\mathcal{Y}| \times |\mathcal{X}| \times |\mathcal{G}|$ rather than with N. The computation is particularly efficient when the full-data model is conjugate.

4.3. Updated Individual Race Probabilities

The EM algorithm produces as a byproduct the updated individual race probabilities $\mathbb{P}(R \mid G, X, Y)$ that condition on Y, unlike the input probabilities $\mathbb{P}(R \mid G, X)$ which do not. Because these updated probabilities condition on Y, the asymptotic bias term in Theorem 3.2 becomes zero, and so it is appropriate to apply the weighting estimator to them to estimate disparities. In fact, the weighting estimate from the updated probabilities is numerically identical to the BIRDiE estimate. While more study is required, for downstream settings where weights are needed, generating these weights with BISG followed by BIRDiE will likely produce more accurate results than simply using BISG weights alone.

4.4. Additional Explanatory Variables

Often, researchers are interested in not just $\mathbb{P}(Y \mid R)$ but also $\mathbb{P}(Y \mid W, R)$, for some variable $W \in \mathcal{W}$ which is not part of the BISG predictors (X, G). For example, a bank auditing potential racial disparities in lending decisions may be interested both in how the rate of loan approval (Y) varies by race, but also

[1] When parameters are estimated as part of the race imputation model, the estimation uncertainty creates dependence in $\widehat{\mathbf{P}}_i$ between observations, which is not accounted for in the BIRDiE posterior. This dependence remains even if, as is standard, the first-stage model parameters and the BIRDiE model parameters are assumed to be a priori independent. In practice, this dependence is minimal compared to the underlying uncertainty in R_i, especially with large datasets where global parameters are estimated precisely, and the effect on downstream analyses is small.

in how loan approval varies by race conditional on a measure of creditworthiness (W). The unconditional disparities reflect factors that may include the realities of systemic racism and inequality, while the conditional disparities measure the fairness of the firm's lending decisions after controlling for these factors. Such estimates could be used to compute various measures of algorithmic fairness. Another scenario is policy evaluation, where researchers are interested in how the impact of policy varies across racial groups. Such an analysis requires incorporating an interaction between race and the treatment variable.

There are two main ways to perform such an analysis with our proposed methodology. The first, and perhaps simplest, is to apply the methodology to the combined variable $\underline{YW} \in \mathcal{Y} \times \mathcal{W}$. This will produce estimates of $\mathbb{P}(Y, W \mid R)$, from which $\mathbb{P}(Y \mid W, R)$ can be straightforwardly calculated by appropriate normalization. This approach will work well if $|\mathcal{Y}|$ and $|\mathcal{W}|$ are both small, so that $|\mathcal{Y} \times \mathcal{W}|$ is of manageable size. If one of these variables has many levels, however, directly estimating the distribution of $Y, W \mid R$ could be less efficient, as it does not account for any structure in the joint (Y, W) distribution.

An alternative approach is to first apply the proposed methodology to estimate $\mathbb{P}(W \mid R)$. This allows for calculation of model-updated BISG probabilities $\tilde{\mathbf{P}}_{|W} = \pi(\mathbf{R} \mid \widehat{\Theta}, \mathbf{W}, \mathbf{G}, \mathbf{X}, \mathbf{S})$, which are also computed as a byproduct of the EM algorithm described above. Then, the methodology can be applied again, using $\tilde{\mathbf{P}}_{|W}$ as the input probabilities rather than the original BISG probabilities, to estimate $\mathbb{P}(Y \mid W, R)$. This approach will likely perform better when W consists of multiple predictors and it may be helpful to shrink $\mathbb{P}(Y \mid W, R)$ towards $\mathbb{P}(Y \mid R)$.

Both approaches require the following assumption, which generalizes Assumption CI-YS.

Assumption CI-YWS (Conditional independence of outcome, predictor and name). For all i, $(Y_i, W_i) \perp\!\!\!\perp S_i \mid R_i, G_i, X_i$, or, equivalently, $W_i \perp\!\!\!\perp S_i \mid R_i, G_i, X_i$ and $Y_i \perp\!\!\!\perp S_i \mid W_i, R_i, G_i, X_i$.

In the lending example, the assumption implies that a measure of creditworthiness is independent of last name after controlling for race, location, and covariates, and that lending decisions are independent of last names after controlling for creditworthiness, race, location, and covariates.

4.5. Addressing Potential Violations of the Assumptions

BIRDiE relies on Assumption CI-YS for identification. In addition, like the weighting and thresholding estimators, it requires Assumptions CI-SG and ACC, which guarantee the accuracy of BISG race probabilities. Unfortunately, these assumptions may not exactly hold in practice, and are also not testable in observed data. In this section and Appendix D, we develop sensitivity analyses that assess how violations of these assumptions affect the estimates of racial disparities.

First, BIRDiE assumes that conditional on unobserved race and observed covariates, outcomes and surnames are independent. As discussed in Section 4.1, however, association between the outcome and country of origin or racial subgroups may lead to correlation between surnames and outcome even after controlling for race and geography. To address this, suppose that a low-dimensional summary statistic of surname, $f : \mathcal{S} \to \mathbb{R}^d$, $d \ll |\mathcal{S}|$, is available, where f maps each surname to a finer ethnic group within each racial category. For example, Imai is a Japanese name whereas McCartan is a name of Irish origin. If f classifies surnames into finer racial subgroups or countries of origin—even approximately—then it can be used to control for this channel of possible violations of Assumption CI-YS. Formally, we relax Assumption CI-YS as follows.

Assumption CI-YSF (Partial conditional independence of outcome and name). For all i,

$$Y_i \perp\!\!\!\perp S_i \mid f(S_i), R_i, G_i, X_i.$$

The next theorem shows that it is still possible to identify racial disparities under Assumption CI-YSF under the identification condition, which is only slightly stronger than for Theorem 4.1.

Theorem 4.4 (Nonparametric Identification Under Assumption CI-YSF). Let $f : \mathcal{S} \to \mathbb{R}^d$, $d < |\mathcal{S}|$, with range $f(\mathcal{S})$. For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, $z \in f(\mathcal{S})$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{R}|}$ with entries $p_{sr} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b} \in \mathbb{R}^{|\mathcal{S}|}$ with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then under Assumption CI-YSF, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, f(S) = z, G = g, X = x)$ are identified if and only if both $\mathbf{P}$ and the augmented matrix $(\mathbf{P} \quad \mathbf{b})$ have rank $|\mathcal{R}|$.

As long as the dimension d of the surname summary statistic $f(S)$ is much smaller than the (usually large) number of surnames $|\mathcal{S}|$, racial disparities are likely to be identified under Theorem 4.4. Thus, Assumption CI-YSF and Theorem 4.4 can be used in conjunction with carefully chosen f in order to probe likely failure modes of the more restrictive Assumption CI-YS. If estimates do not change by a substantively large amount when $f(S)$ is included, then researchers can be more confident in the plausibility of Assumption CI-YS. We demonstrate this approach in Section 5.4.

Second, bias can also arise from violations of the assumptions underlying the BISG methodology (Assumptions CI-SG and ACC). Of course, this is not unique to the proposed methodology: violations of these assumptions will also affect the validity of other disparity estimators such as weighting or thresholding. However, since as discussed above the BISG assumptions may rarely hold exactly in practice, we provide in Appendix D several results characterizing how the model's estimates are affected by bias in the BISG probabilities.

5. Empirical Validation with the Voter File

To better understand how BIRDiE performs in real-world contexts, we apply it to North Carolina voter registration data. Since this data contains individual-level self-reported race for almost all voters, the "ground truth" relationship between outcome and race is known for this subset. We compare the performance of BIRDiE models and the OLS estimator against those of the weighting and thresholding estimators. We also evaluate how the estimation error depends on the geographic level used in

the BISG probabilities. Finally, we demonstrate a diagnostic for potential violations of the identifying assumption (Section 4.5). Appendix C contains further discussion of data and models, along with additional study of the OLS estimator and extensions of the BIRDiE methodology to small-area estimates (Section 4.2), improved individual race predictions (Section 4.3), and estimation conditional on an additional explanatory variable (Section 4.4).

5.1. North Carolina Voter File

Like most other Southern states, which have a history of disenfranchising minority voters, the state of North Carolina (NC) asks (and previously required) every voter to self-report their race upon registration. This data, along with voters' names, addresses, gender, party registration (if any), and voting history, is part of the voter file that the secretary of state makes publicly available. This feature makes the voter file an ideal validation setting. The outcome we examine here, party registration, is the product of many unobservable factors, and is known to differ across racial groups. Since self-reported race is available, inferences about these racial disparities using the estimators discussed here can be compared to the corresponding ground truth.

Estimation of party registration by race is of substantive interest as well, especially in the context of the Voting Rights Act of 1965 (VRA). The relationship between these variables is critical for understanding the impact of policy changes such as redistricting or election rules on compliance with the VRA, and for establishing legal standing to challenge these policies under the VRA. As many states do not ask for self-reported race during voter registration, methods like BIRDiE are important tools for evaluating VRA compliance.

We use a subset of the October 2022 voter file which could be linked to a proprietary voter file provided by L2, Inc., a leading national non-Partisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters, and consultants for use in campaigns. The L2 file geocoded each address to a Census block, which allows for the finest block-level BISG predictions. We also removed any records without individual race information, since our goal is validation compared to some ground truth, rather than inference about the entire population of registered NC voters. Altogether, 22.1% of records either had missing race information or could not be linked to the L2 file.

5.2. The Model Setup

We first calculate BISG probabilities using 2010 Census data at the census block, tract, ZIP code tabulation area (ZCTA), and county level. Every record in the voter file contains county information, while roughly 13% of records are missing ZIP codes and 27% of records are missing blocks/tracts; when these finer geographic identifiers were missing, we used county-level Census tables in the BISG calculations.

The BISG probabilities are broadly accurate. Using the maximum a posteriori racial category as a prediction, we obtain accuracy of 76.2% for the county probabilities, 78.4% for the ZIP code probabilities, 78.5% for the tract probabilities, and 79.6% for the block probabilities.[2]

Since the goal of our validation study is to compare BIRDiE estimates with weighting and thresholding estimates, we do not make additional comparisons between BISG probabilities and those generated with alternative racial prediction methods. To the extent competing racial prediction methods improve prediction accuracy, we expect the gap between different disparity estimation methods (weighting, thresholding, BIRDiE) to narrow, consistent with Theorems 3.2 and 4.3. As we have discussed, however, high accuracy of racial prediction alone is neither necessary nor sufficient for accurate estimation of racial disparities. If other racial prediction methods produce increased accuracy at the cost of worsened calibration, accuracy in estimating racial disparities may be poor whether using weighting, thresholding, or BIRDiE estimates.

In our validation, for a given set of BISG probabilities, we estimate the conditional distribution of each outcome variable given race using BIRDiE with both saturated pooling and multinomial mixed-effects models described in Appendix B.1. We then compare the resulting estimates based on these BIRDiE models against those of the two existing estimators—the weighting estimator as well as a thresholding estimator that deterministically assigns each individual the maximum a posteriori racial category. We also compare the results to those obtained by the OLS estimator described in Section 4.1. To give an idea of sampling variability, we fit the saturated BIRDiE model using the Gibbs sampler described above, run for 500 post-warmup iterations. We also bootstrap the estimates for the weighting and threshold estimators, and aggregate the OLS standard errors through the poststratification process. It was not computationally feasible to bootstrap the mixed-effect model nor perform full MCMC to obtain posterior samples.

5.3. Estimates of Racial Disparity in Party Registration

We first examine the relative accuracy of the proposed methods in estimating the disparity between White and Black, and White and Hispanic voters, in party registration. For example, the true difference in Democratic registration between Black and White voters in the sample is 54.6 percentage points (pp), meaning Black voters register Democratic at a much higher rate. However, the standard weighting approach produces an estimate of only 16.8 pp for this disparity—less than half the true value. This is consistent with Corollary 3.2.1, which states that the weighting estimator tends to underestimate the magnitude of racial disparity. The thresholding estimator, while slightly better, also misses the mark, with an estimate of 26.5 pp. In contrast, the saturated BIRDiE model produces an estimate of 48.9 pp, and the mixed BIRDiE model also estimates 48.8 pp. These estimates are only slightly lower than the ground truth.

Figure 2 compares the empirical performance of the BIRDiE models against that of the weighting and thresholding estimators across all of these possible disparity measurements, using the county-level BISG predictions. For White–Black (left plot) and White-Hispanic (right plot) disparities in party registration, the BIRDiE models (solid circles and squares) and the OLS esti-

[2]That is, we measure the fraction of the time the most-likely racial category agrees with the true racial category.

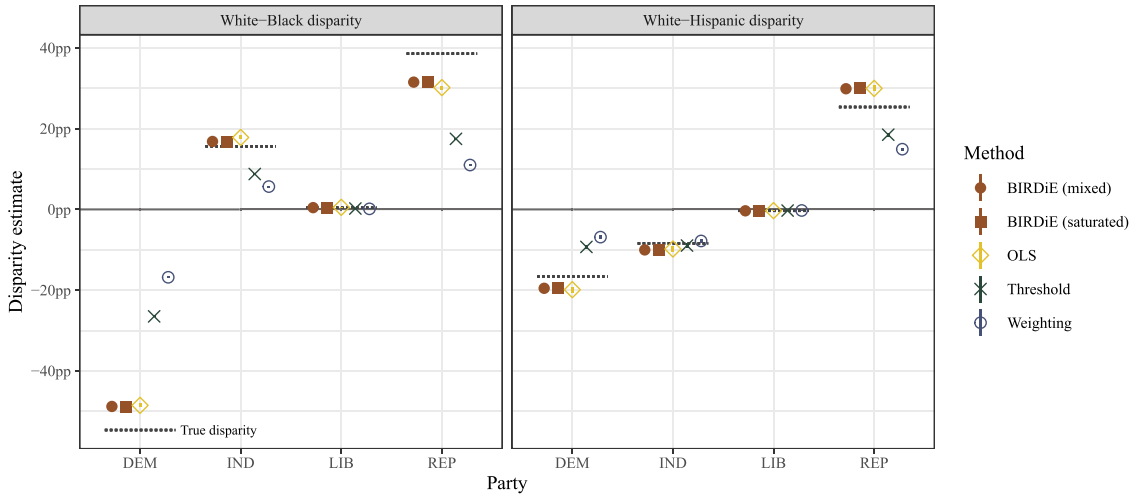


Figure 2. Error in the White-Black and White-Hispanic disparity estimates for party registration, by estimation method. The true disparities are indicated by the dotted lines. All methods used county-level data for this figure; the results for other levels of geographic detail are generally similar, except for the OLS estimator. Estimation uncertainty is shown as a 95% confidence interval for all methods other than the mixed model, but is generally too small to be seen.

mator substantially outperform the two commonly used estimators (open circles and crosses). For two major parties, both the weighting and thresholding estimators exhibit a substantial amount of estimation error, for example, exceeding 20 pp for the White-Black disparity for the Democratic party. In contrast, the two BIRDiE models yield a much smaller estimation error that ranges within several percentage points for all racial disparity estimates. The saturated and mixed-effects BIRDiE models perform similarly with no discernible difference, while the OLS estimator is more variable and performs very slightly worse for the White–Black disparity estimates.

Section C contains additional validation that compares the total variation distance between the estimated joint distribution of party and race and the actual joint distribution.

5.4. Conditional Independence Diagnostic

Though the results above show that BIRDiE improves considerably on existing estimation approaches, the agreement with the ground truth is not perfect. Consequently, we examine the sensitivity of our party registration estimates to potential violations of Assumption CI-YS, following the method outlined in Section 4.5 that is based on a low-dimensional summary statistic of surnames. Our statistic is based on a publicly available sample of 5% of the individual records for the 1930 Census (Ruggles et al. 2021), which contains individual names, individual and parental birthplace, and detailed race, ethnicity, and tribal codes. Since many regions of Asia, particularly Vietnam, experienced little emigration to the United States before 1930, we further supplement this data with around 3000 Asian surnames classified into six regional subgroups: Chinese, Filipino, Indian, Japanese, Korean, Vietnamese, NHPI, and Other (Lin et al. 2025).

Using these subgroups and the 1930 birthplace and racial data, we can classify most surnames in the voter file into nine groups (see Appendix E for a brief description of the groupings and the most common 50 surnames for each group). While somewhat arbitrary, these groups are chosen to combine countries of origin which had significant immigration to the U.S. during similar periods.

We first evaluate the plausibility of Assumption CI-YS by examining the correlation between the residuals of the BIRDiE model fit and indicator variables for each of the nine surname groups. Under Assumption CI-YS, this residual correlation should be zero everywhere. As Figure 3 shows, however, for many groups and party labels, the correlation is small but deviates from zero more than would be expected given only sampling variation. Here, we use the residuals from the county-level saturated model specification, but the results are not sensitive to this choice.

Notably, voters with names in the Anglosphere and Black surname group, which includes surnames that are relatively more common among many-generation residents of the U.S., such as Smith, Williams, and Brown, are significantly less likely to register as Democrats and independents, and more likely to identify as Republicans, even after controlling for race and geography. Meanwhile, voters with names in the First and Second wave European immigration surname groups, which include surnames more common among 19th and 20th century immigrants from Europe, display the opposite pattern. Differences among surname groups designed to correlate with membership in various Asian subgroups are also visible. As might be expected, the relatively many significant correlations are indicative more of the large number of observations in the data, rather than large residual correlations themselves—all of the correlations are quite small in magnitude, with most on the order of 0.01 or so. Thus, we might expect our party-by-race estimates to be little affected by the inclusion of the surname group indicators. In other words, the violations of Assumption CI-YS are statistically significant, but may not be substantively so.

Indeed, re-fitting the county-level saturated model with an additional surname group covariate produces nearly identical results, albeit at a moderately higher computational cost given the increased number of $(G, X, f(S))$ cells. This re-fitting requires Assumption CI-YSF, which relaxes Assumption CI-YS. We find that the average party registration rate estimate changes only by 1.0 pp, with the largest change being 3.6 pp, for the rate of Republican registration among Other voters. These changes are small compared to the underlying disparities,

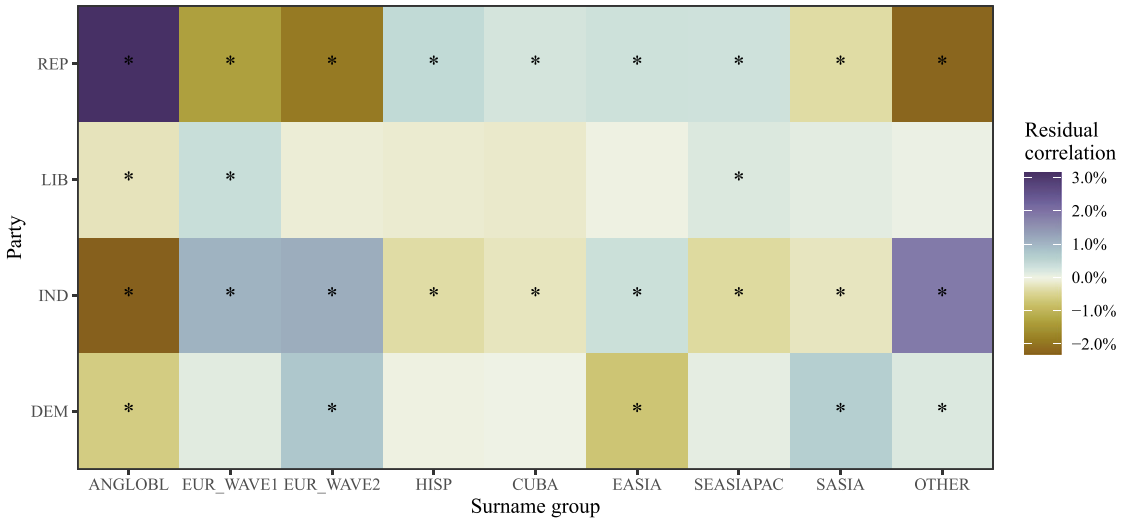


Figure 3. Residual correlation between party registration and nine surname groups, after controlling for race and location. Correlations whose 90% Wishart confidence intervals exclude zero are marked with an asterisk. These confidence intervals do not account for the dependence in the residuals due to the model fitting, and thus are likely anti-conservative. See Appendix E for details on the surname groups.

which are on the order of 10–40 pp. The accuracy of the updated BISG probabilities is likewise virtually unchanged. All in all, this analysis provides confidence that violations of Assumption CI-YS for the NC voter file are likely minor and would have minimal effects on our findings. Of course, if more precise estimates of party registration were required, then the changes might be considered sufficiently large to warrant more careful consideration of possible $f(S)$ that could remedy the violation.

6. Analysis of Tax Data

6.1. Estimation Procedure

We use a random 10% sample of individual tax returns (Form 1040s) filed with the IRS for tax year 2019, a total of 17,145,898 observations. To calculate individual race probabilities for every observation, we use the ZIP code tabulation area (ZCTA) corresponding to the geocoded address listed on the return, plus the last name of the primary filer using a standard BISG model. This means that conclusions about racial groups here refer to the race of the primary filer, and not the race of other household members. For the roughly 3.4 million records for which geocoding was not successful, only last names were used.

The outcome variable is the amount of the HMID claimed by the filer, discretized into 11 levels: one for a deduction of $0, capturing roughly 90% of the sample, and ten levels corresponding to the deciles of the HMID among those taking the deduction. Given the size of the data, our outcome model is the no-pooling model for HMID level by geography and racial group. We coarsen the geography variable used for modeling to the Public Use Microdata Area (PUMA) level.

Further modeling and estimation details may be found in Appendix F.

6.2. Findings

Figure 4 shows the estimated proportion of filers in each racial group which take the HMID at all, and the distribution of the HMID claim amount among those who do. Racial disparities are immediately apparent: while 10.6% of White filers take the HMID, just 6.5% of Black filers and 4.6% of Hispanic filers do. In contrast, roughly 12% of Asian filers take the deduction.

How much of this disparity is explained by differences in home ownership rates across racial groups? We develop estimates of the fraction of each racial group that has a mortgage for their home based on the 2010 decennial census (see Appendix F.2 for details). We then plot as dashed lines in Figure 4(a) the estimated fraction of each racial group that would claim the HMID if every filer with a mortgage claimed the deduction at the same rate as White filers do. The figure shows that the lower share of Black taxpayers claiming the HMID can be explained by the lower home ownership rate among that group. However, for other groups, disparities remain after controlling for having a mortgage. In particular, Hispanic filers claim the HMID at a 2.5 percentage point lower rate than would be expected based on mortgage rates alone. Asian filers claim the HMID at a higher rate than their share of the population with mortgages would imply. These results suggest that closing disparities in home ownership may not be sufficient to eliminate disparities in who benefits from the HMID, as evidently other aspects of filers' situations beyond home ownership, such as eligibility for other itemized deductions, are affecting whether their HMID benefits.

Beyond disparities in the rate that taxpayers claim any HMID, there are also racial differences in the amount of the HMID among claimants. This is apparent in both the distribution across HMID deciles in Figure 4(a) as well as in Figure 4(b), which displays estimates of the mean HMID amount among filers who claim the deduction. These estimates were produced by weighting the observed HMID amounts according to BIRDiE-updated race probabilities, as described in Section 4.3. Compared to White claimants, whose average deduction is $13,500, HMID amounts for Black claimants are skewed toward the lower deciles, translating to a $2100 to $3800 lower average HMID amount for these groups. In contrast, Hispanic claimants deduct just $500 less than White claimants, and Asian claimants deduct

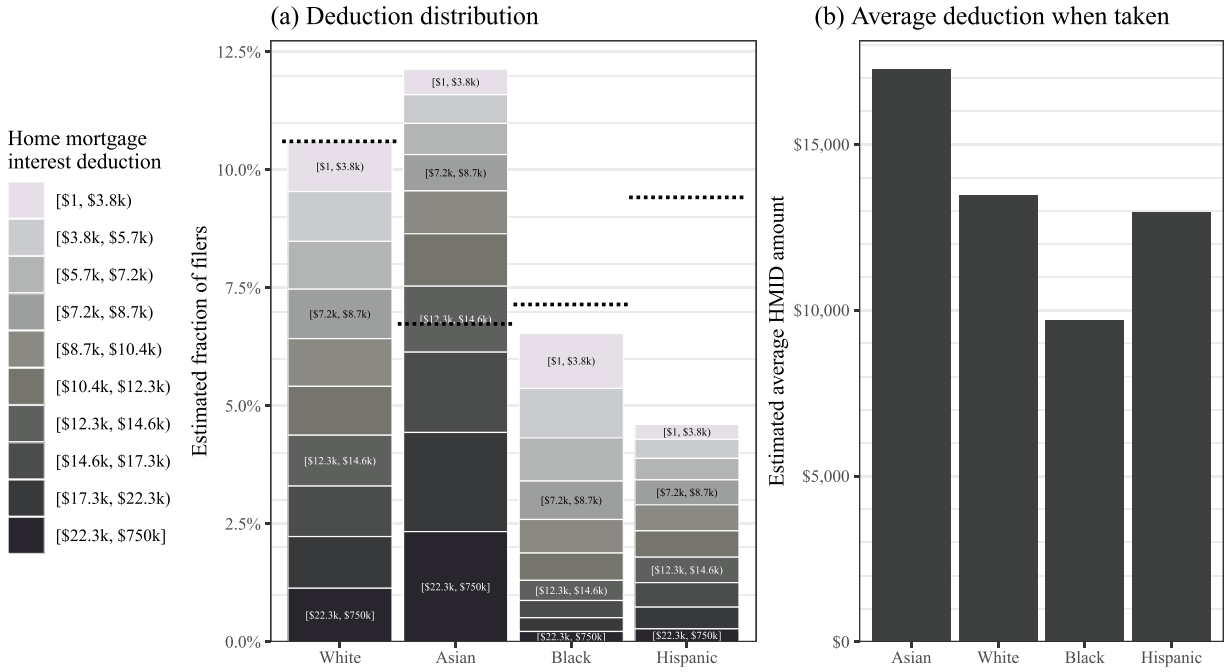


Figure 4. Estimates of usage of the home mortgage interest deduction by race. Panel (a) shows estimates of the proportion of filers who took the deduction, further broken into deciles by deduction amount. Panel (b) shows estimates of the average size of the deduction among filers who took the deduction at all. The dotted lines mark the expected height of each bar if the only disparities were those in mortgage rates between racial groups.

Table 1. Estimates of HMID claim rate and average deduction by racial group.

Average claim	White	Black	Hispanic	Asian
Rate	10.6%	6.5%	4.6%	12%
Among claimants	$13,500	$9,700	$13,000	$17,300
Unconditionally	$1,400	$600	$600	$2,100

$3800 more, on average. Asian filers' deductions fall into the highest deciles at more than double the rate of any other group. In fact, a higher fraction of Asian filers take at least a $17,000 deduction than the fraction of Hispanic filers who take any deduction at all. Table 1 compares these estimates for claimants to the unconditional averages of the HMID benefit amount across racial groups.

Overall, our findings support the claims of researchers such as Moran and Whitford (1996) and Brown (2022) that the HMID is disproportionately unavailable to Black and Hispanic taxpayers. Our estimates show that the picture is complicated further by differences between racial groups even accounting for the prevalence of mortgages. In addition, the pattern of disparities in overall HMID claims looks different from the disparities in the amount of the HMID among claimants.

7. Discussion

We have introduced a new identifying assumption and accompanying model, BIRDiE, and clarified other assumptions implicit in approaches to disparity estimation when individual race is not observed. In many real-world applications, we believe that the new model and identification condition are appropriate and will produce significantly improved estimates. However, there is no one-size-fits-all approach for the estimation of racial disparities. For example, the existence of name-based discrimination may violate our identification assumption especially when racial categories, for which data are available, are coarse. Although we provide a diagnostic that partially addresses this concern for a likely violation pathway, careful consideration of the underlying causal and information structure is required to avoid making the incorrect conclusions.

As our empirical studies show, in realistic settings BIRDiE can substantially outperform existing estimators of racial disparities, both in aggregate and for small areas. Albright and Gamboa-Arbelaez (2024) also conducts an empirical validation of various race imputation methods, finding that BIRDiE performs better than other methods. The BIRDiE methodology also produces improved BISG probabilities, and can be used to estimate disparities conditional on other variables. These additional features should prove helpful in practical settings. Appendix G contains additional discussion of these points, including practical recommendations for users of BIRDiE and ethical considerations when applying methods like BISG and BIRDiE.

Supplementary Materials

Supplementary material contains proofs of all propositions, details on BIRDiE models and computation, additional discussion and results for the validation analysis, proposed sensitivity analyses, further details on the tax study, and additional discussions and recommendations for practitioners.

Acknowledgments

We thank Bruce Willsie, CEO of L2, Inc., for providing us with the geocoded voter file we use in this article. We also thank Hiroto Katsumata, Soichiro Yamauchi, and an anonymous reviewer of the Alexander and Diviya Magaro Peer Pre-Review Program for useful feedback. The views expressed in this article are those of the authors and do not necessarily represent the views of the US Treasury Department. Any taxpayer data used in this research was kept in a secured IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed.

ORCID

Cory McCartan http://orcid.org/0000-0002-6251-669X
Jacob Goldin http://orcid.org/0000-0001-5518-0027
Daniel E. Ho http://orcid.org/0000-0002-2195-5469
Kosuke Imai http://orcid.org/0000-0002-2748-1022

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Data Availability Statement

Replication code and data are available at *https://github.com/CoryMcCartan/birdie-replication*.

References

Albright, A., and Gamboa-Arbelaez, J. (2024), "Imputing Race," working paper. [2151]

Anderson, M., and Fienberg, S. E. (1999), *Who Counts?: The Politics of Census-Taking in Contemporary America*, New York: Russell Sage Foundation. [2142]

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2024), "PPI++: Efficient Prediction-Powered Inference," arXiv preprint arXiv:2311.01453. [2141]

Angrist, J. D., and Krueger, A. B. (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American statistical Association*, 87, 328–336. [2145]

Argyle, L., and Barber, M. (2024), "Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records." *American Political Science Review*, 118, 1058–1066. DOI:10.1017/S0003055423000229. [2140,2141,2143,2144,2146]

Åslund, O., and Skans, O. N. (2012), "Do Anonymous Job Application Procedures Level the Playing Field?" *ILR Review*, 65, 82–107. [2144]

Brown, D. A. (2022), *The Whiteness of Wealth: How the Tax System Impoverishes Black Americans–and How We Can Fix It*, New York: Crown. [2140,2141,2151]

Budiman, A., Cilluffo, A., and Ruiz, N. G. (2019), *Key Facts About Asian Origin Groups in the US*, Washington, DC: Pew Research Center. [2143]

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019), "Fairness Under Unawareness: Assessing Disparity When Protected Class is Unobserved," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348. [2141,2143]

Cheng, L., Gallegos, I. O., Ouyang, D., Goldin, J., and Ho, D. (2023), "How Redundant Are Redundant Encodings? Blindness in the Wild and Racial Disparity When Race is Unobserved," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 667–686. [2143]

Cho, W. T., and Manski, C. F. (2008), "Cross-Level/Ecological Inference," in *Oxford Handbook of Political Methodology*, pp. 547–569, Oxford: Oxford University Press. [2140]

Congressional Research Service. (2017), "Tax Deductions for Individuals: A Summary," Technical Report R42872. [2141]

Cronin, J. A., DeFilippes, P., and Fisher, R. (2023), "Tax Expenditures by Race and Hispanic Ethnicity: An Application of the U.S. Treasury Department's Race and Hispanic Ethnicity Imputation," Technical Report 122, Office of Tax Analysis, U.S. Department of the Treasury. [2142]

Crossley, T. F., Levell, P., and Poupakis, S. (2022), "Regression with An Imputed Dependent Variable," *Journal of Applied Econometrics*, 37, 1277–1294. [2145]

Decter-Frain, A. (2022), "How Should We Proxy for Race/Ethnicity? Comparing Bayesian Improved Surname Geocoding to Machine Learning Methods," arXiv:2206.14583. [2140,2143,2144,2146]

DeLuca, K., and Curiel, J. A. (2022), "Validating the Applicability of Bayesian Inference with Surname and Geocoding to Congressional Redistricting," *Political Analysis*, 31, 465–471. [2140,2143]

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (2024), "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses," Working paper. Available at https://naokiegami.com/paper/dsl_ss.pdf. [2141]

Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. (2008), "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity," *Health Services Research*, 43, 1722–1736. [2140,2142]

Elzayn, H., Smith, E., Hertz, T., Ramesh, A., Fisher, R., Ho, D. E., and Goldin, J. (2023), "Measuring and Mitigating Racial Disparities in Tax Audits." [2141]

Fiscella, K., and Fremont, A. M. (2006), "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity," *Health Services Research*, 41, 1482–1500. [2140,2142]

Fong, C., and Tyler, M. (2021), "Machine Learning Predictions as Regression Covariates," *Political Analysis*, 29, 467–484. [2141]

Goodman, L. A. (1953), "Ecological Regressions and Behavior of Individuals," *American Sociological Review*, 18, 663–664. [2140]

Greengard, P., and Gelman, A. (2023), "BISG: When Inferring Race or Ethnicity, Does It Matter that People Often Live Near their Relatives?" [2140,2142,2143,2144]

Greenwald, D., Howell, S. T., Li, C., and Yimfor, E. (2023), "Regulatory Arbitrage or Random Errors? Implications of Race Prediction Algorithms in Fair Lending Analysis," Technical Report, National Bureau of Economic Research. [2141]

Imai, K., and Khanna, K. (2016), "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records," *Political Analysis*, 24, 263–272. [2140,2143]

Imai, K., Lu, Y., and Strauss, A. (2008), "Bayesian and Likelihood Inference for 2× 2 Ecological Tables: An Incomplete-Data Approach," *Political Analysis*, 16, 41–69. [2140]

Imai, K., Olivella, S., and Rosenman, E. T. (2022), "Addressing Census Data Problems in Race Imputation via Fully Bayesian Improved Surname Geocoding and Name Supplements," *Science Advances*, 8, 1–10. [2140,2142,2143,2144,2146]

Kallus, N., Mao, X., and Zhou, A. (2022), "Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination," *Management Science*, 68, 1591–2376. [2141]

Kenny, C., McCartan, C., Kuriwaki, S., Simko, T., and Imai, K. (2024), "Evaluating Bias and Noise Induced by the U.S. Census Bureau's Privacy Protection Methods," *Science Advances*, 10, eadl2524. [2142]

Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E. T., Simko, T., and Imai, K. (2021), "The Use of Differential Privacy for Census Data and Its Impact on Redistricting: The Case of the 2020 US Census," *Science Advances*, 7, eabk3283. [2140,2143]

King, G. (1997), *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*, Princeton: Princeton University Press. [2140]

Knox, D., Lucas, C., and Cho, W. K. T. (2022), "Testing Causal Theories with Learned Proxies," *Annual Review of Political Science*, 25, 419–441. [2145]

Kuroki, M., and Pearl, J. (2014), "Measurement Bias and Effect Restoration in Causal Inference," *Biometrika*, 101, 423–437. [2145]

Lin, Q., Ouyang, D., Guage, C., Gallegos, I., Goldin, J., and Ho, D. (2025), "Enabling Disaggregation of Asian American Subgroups: A Dataset of Wikidata Names for Disparity Estimation," *Scientific Data* 12, 580. [2149]

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018), "Identifying Causal Effects with Proxy Variables of An Unmeasured Confounder," *Biometrika*, 105, 987–993. [2145]

Moran, B. I., and Whitford, W. (1996), "A Black Critique of the Internal Revenue Code," *Wis. L. REv.*, p. 751. [2140,2151]

Park, J., Malachi, E., Sternin, O., and Tevet, R. (2009), "Subtle Bias Against Muslim Job Applicants in Personnel Decisions," *Journal of Applied Social Psychology*, 39, 2174–2190. [2144]

Rosenman, E. T., Olivella, S., and Imai, K. (2023), "Race and Ethnicity Data for First, Middle, and Last Names," *Scientific Data*, 10, 1–11. [2142]

Ruggles, S., Fitch, C. A., Goeken, R., Hacker, J. D., Nelson, M. A., Roberts, E., Schouweiler, M., and Sobek, M. (2021), "Ipums Ancestry Full Count Data: 1930 5version 3.0." [2149]

Selén, J. (1986), "Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical Data," *Journal of the American Statistical Association*, 81, 75–81. [2141]

Strmic-Pawl, H. V., Jackson, B. A., and Garner, S. (2018), "Race Counts: Racial and Ethnic Data on the U.S. Census and the Implications for Tracking Inequality," *Sociology of Race and Ethnicity*, 4, 1–13. [2142]

Sullivan, L., Meschede, T., Shapiro, T., and Fernanda Escobar, M. (2017), "Misdirected Investments: How the Mortgage Interest Deduction Drives Inequality and the Racial Wealth Gap," Technical Report, Institute on Assets and Social Policy and the National Low Income Housing Coalition. [2142]

U.S. Census Bureau. (2014), "Frequently Occurring Surnames from the 2010 Census," available at *https://www.census.gov/topics/population/genealogy/data/2010_surnames.html*. Accessed: 2022-03-25. [2142]

——— (2022), "Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census," March 10, 2022, Release Number CB22-CN.02. [2142]

Voicu, I. (2018), "Using First Name Information to Improve Race and Ethnicity Classification," *Statistics and Public Policy*, 5, 1–13. [2140]

Wakefield, J. (2004), "Ecological Inference for 2× 2 Tables," *Journal of the Royal Statistical Society*, Series A, 167, 385–425. [2140]

Zest AI. (2020), "Zest Race Predictor (zrp)," available at *https://github.com/zestai/zrp/*. [2140,2143,2144,2146]

Zhang, Y. (2018), "Assessing Fair Lending Risks Using Race/Ethnicity Proxies," *Management Science*, 64, 178–197. [2143]