# Imputing Race

Alex Albright*
Juliana Gamboa-Arbelaez†

October 2025

**Abstract**

Missing data on racial identity is a fundamental challenge to studying race in the economy. When race is unobserved, researchers often impute it from variables such as names and locations. Although many imputation methods exist, systematic comparisons of their performance are rare. This paper offers one of the first such comparisons, evaluating leading methods across multiple dimensions. No single method dominates – the best choice depends on the object of interest. These findings highlight the limitations of a common empirical practice: choosing methods based on a single performance dimension.

*Albright: Federal Reserve Bank of Minneapolis, Opportunity and Inclusive Growth Institute (Email: alex@albrightalex.com)

†Gamboa-Arbelaez: University of Minnesota, Economics Department (Email: gambo038@umn.edu)

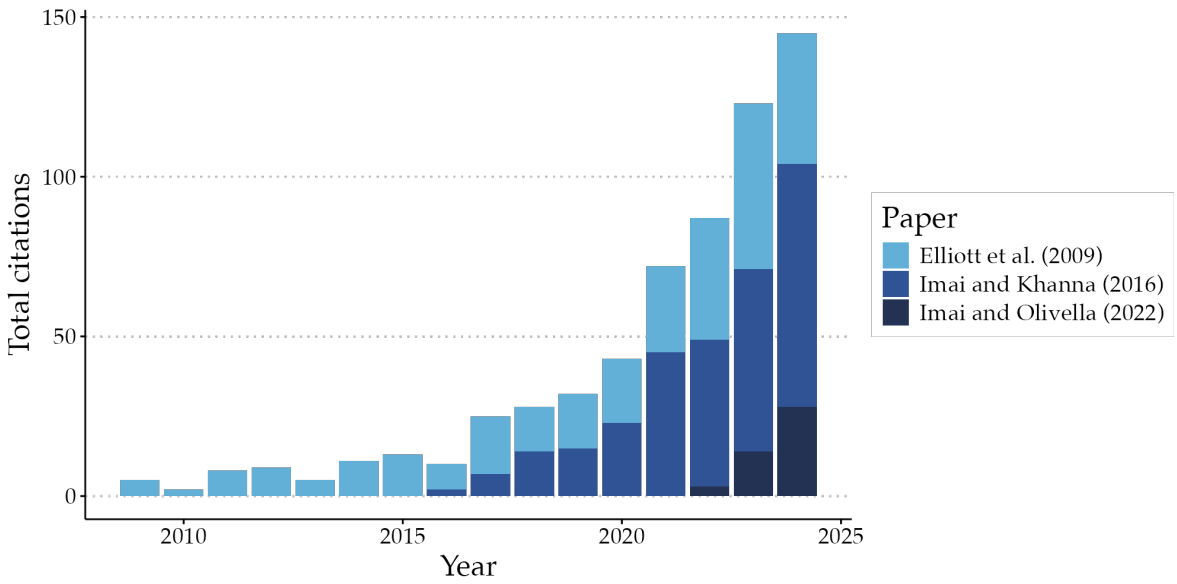
We thank Abbie Wozniak, Chi Hyun Kim, Jacob Goldin, Tomás Monarrez, and participants at the Federal Reserve System Economic Heterogeneity Conference for useful feedback on this project. This project would not be possible without the contributions of many people who develop and maintain the open source tools we use in this paper. The views expressed in this paper are our own and do not necessarily represent those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

# 1 Introduction

Studying economic outcomes across racial groups is crucial for understanding economic inequality in the United States. However, many datasets lack information on race, which creates an obvious challenge for empirical work. How can researchers study racial heterogeneity when race itself is unobserved? In practice, researchers increasingly *impute* race, meaning they predict racial identity using other observables, such as name and geographic location.

Imputing race has enabled recent research in economics on many topics, including taxes (Elzayn et al., 2025), housing (Diamond and Diamond, 2024), lending and credit (Frame et al., 2021; Howell et al., 2024), and the justice system (Feigenberg and Miller, 2022).[1] Research using race imputation is rapidly accelerating: citations to Bayesian Improved Surname Geocoding (BISG), a widely used method, more than tripled between 2020 and 2024 (Figure 1).

Figure 1: Growing Citations to a Common Race Imputation Method




*Notes:* This figure shows the total google scholar citations over time to Bayesian Improved Surname Geocoding (BISG), a commonly used race imputation method. BISG was developed and improved upon in three papers: Elliott et al. (2009), Imai and Khanna (2016), and Imai et al. (2022). The figure is a stacked bar chat which splits citations out across the three papers.

The growing prevalence of race imputation reflects both the expanding availability of methods and shifts in the types of data used in empirical research. In economics, researchers

[1] Race imputation has also been used to study the opioid crisis (Cohle and Ortega, 2023), eviction (Hepburn et al., 2020), policing (Edwards et al., 2019), and COVID (Labgold et al., 2021).

increasingly rely on administrative and private datasets (Goldsmith-Pinkham, 2024; Einav and Levin, 2014).[2] These types of datasets often lack direct information on race but contain individual-level variables suitable for imputation. Administrative data examples include tax records (Fisher, 2023), healthcare records (Fremont et al., 2016), and court records (Hepburn et al., 2020), while private data examples include credit bureau data (Albanesi and Vamossy, 2024) and proprietary housing data (Box-Couillard and Christensen, 2024).

Although imputing race underlies a growing body of research, systematic comparisons of imputation methods remain limited. How do researchers impute race, how well do these methods perform, and on which dimensions? The lack of broad comparisons is understandable: methods are proliferating, implementations are fragmented across software environments, and research incentives often favor developing new methods over evaluating existing ones (Christensen and Miguel, 2018; King, 1995). Yet addressing these questions is crucial, because the choice of imputation method can have first-order consequences for empirical conclusions.

Our paper addresses this gap in the literature by conducting a comprehensive comparative analysis on race imputation methods. To enable comparison, we focus on methods that can be applied using publicly available US data that also include self-reported race – specifically, voter file data. As a result, we compare imputation methods that rely on names, geographic information, or both, since these are the voter file variables commonly used for imputation.[3] We compare standard methods (surname-based imputation, geography-based imputation, and BISG) as well as recently developed methods (Zest Race Prediction, NamePrism, and BIRDiE) that fit this criterion.

When evaluating performance, we distinguish between three distinct performance dimensions: individual-level classification, racial disparity estimation, and population-level racial demographic estimation. Prior work has typically discussed these dimensions separately. Papers that introduce new methods often emphasize improvements in individual-level classification performance (Ye et al., 2017; Chaturvedi and Chaturvedi, 2024; Chalavadi et al., 2025; Li, 2023), while a growing body of work highlights the importance of imputation methods for estimating racial disparities (Greenwald et al., 2024; McCartan et al., 2025). A few papers have discussed two of these three dimensions simultaneously, specifically when comparing standard BISG to alternative machine learning-based methods

[2]Administrative data now appears in 25% of NBER working papers, up from just 5% in 2005 (Goldsmith-Pinkham, 2024). Similarly, 20% of *American Economic Review* articles in 2014 received exemptions from the journal's data availability policy due to private data use, compared to 4% in 2006 (Einav and Levin, 2014).

[3]Race can be imputed using other kinds of data as well, such as images (Greenwald et al., 2024; Cook et al., 2025; Hurtado and Sakong, 2023) or online behavior (Matthew et al., 2024).

(Chaturvedi and Chaturvedi, 2024; Decter-Frain, 2022; Curiel and DeLuca, 2024a). Our paper is distinctive in evaluating a range of both common and newly developed methods methods across all three dimensions.

Consistent with prior work, we find standard race imputation methods (surname-based imputation, geography-based imputation, and BISG) can perform quite poorly for performance dimensions of interest (Greenwald et al., 2024; McCartan et al., 2025). All methods systemically overcount White people relative to other racial groups. Moreover, many methods underestimate the magnitude of racial gaps by over 50%, and some even reverse the direction of racial gaps.

However, our results also demonstrate that two new methods make notable progress on the shortcomings of standard methods. First, Zest Race Prediction (ZRP) (Matthews et al., 2024), outperforms other direct imputation methods on individual-level classification and describing racial demographics. ZRP also outperforms other direct imputation methods for some racial gaps, but not all. The second new method that makes progress is BIRDiE (McCartan et al., 2025), and is a correction method for estimating racial disparities. Consistent with McCartan et al. (2025), we find that BIRDiE yields very accurate measures of racial disparities, and does so consistently across different racial group combinations.

We also uncover shortcomings with new methods that are gaining steam in applied research. The NamePrism method (Ye et al., 2017), which has been cited by over 150 papers as of August 2025, does not consistently outperform basic surname-based imputation despite its use of proprietary data and more complicated methods. We find evidence that this is driven by its subpar performance in imputing race for Black people. Therefore, newer methods leveraging machine learning are not always an improvement over existing methods.

This paper offers practical comparisons for researchers imputing race, but it also makes broader methodological contributions. Most crucially, our results show that imputation methods can outperform in some dimensions of performance while underperforming in others. If a method outperforms in terms of individual-level classification, this does not guarantee better estimation of racial demographics or racial disparities. These findings highlight the limitations of a common empirical practice: choosing methods based on individual-level performance. Instead, practitioners and researchers must consider their end goal when imputing race, and choose empirical tools based on their research focus (e.g., using BIRDiE (McCartan et al., 2025) for racial gap estimation).

Our paper contributes to several literatures. First, we directly contribute to research on

race imputation methods by providing a comparative analysis of both standard and newly developed approaches using publicly available data. This work builds on early studies of geography-based and surname-based methods (Fiscella and Fremont, 2006) and the development of BISG (Elliott et al., 2009), while extending recent evaluations of machine-learning–based imputation (Curiel and DeLuca, 2024b; Ye et al., 2017) and racial disparity adjustment methods (McCartan et al., 2025).

Second, we contribute to the broader study of race measurement and its implications for empirical research. How race is coded affects public understanding of US demographic change (Starr and Pao, 2024; Ventura and Flores, 2025) and can lead to differences between administrative records and self-reports, with important consequences for estimating racial disparities (Finlay et al., 2024; Baron et al., 2024; Jarrín et al., 2020). Moreover, whether individuals report their race is itself correlated with economic outcomes (García and Darity Jr, 2022).

Finally, our paper informs research on racial inequality in the US economy from a methodological perspective. Race imputation has been used to study a wide range of economic topics, including lending and credit (Consumer Financial Protection Bureau, 2014; Hurtado and Sakong, 2023), housing (Diamond and Diamond, 2024), tax policy (Cronin et al., 2023; Costello et al., 2024), and the justice system (Luh, 2023; Goncalves and Mello, 2021). Our results highlight the underlying importance of race imputation methods in economics research. Studying these methods is increasingly necessary in an era where the proliferation of individual-level data is expanding the possibilities for imputation.

Our paper proceeds as follows. Section 2 provides background on the measurement of race in the US. Section 3 then describes the voter file data we use. Section 4 outlines the different race imputation methods we compare, while Section 5 describes the three different dimensions of performance we study. Section 6 presents the performance results across all methods and dimensions. Section 7 concludes.

# 2 Background

## 2.1 History of Measuring Race in the US

Data on race in the US has been collected since the very first decennial Census in 1790 (Jensen and Jones, 2024). That Census was administered by enumerators and included separate columns for "free whites," "all other free persons," and "slaves" (Jensen and Jones, 2024; Medina and Lai, 2023). In contrast, the most recent decennial Census (2020) asked

respondents to self-identify their race (White; Black or African American; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; and Some Other Race) and their ethnicity (Hispanic or Latino; Not Hispanic or Latino) (Jensen and Jones, 2024; Medina and Lai, 2023).

Over the course of 230 years, there have been dozens of changes to the definitions of race and ethnicity in the Census (Jensen and Jones, 2024). These changes were often driven by shifting political climates, demographic patterns, and significant historical events. For example, in 1870, following increased immigration from China, "Chinese" was added as a distinct race category (Jensen and Jones, 2024). In 1930, Census instructions incorporated a "one drop rule" for racial classification (Cohn, 2010; Jensen and Jones, 2024). The Hispanic ethnicity question was introduced in 1970 (Jensen and Jones, 2024), and beginning in 2000, respondents gained the ability to select multiple racial categories (Jensen and Jones, 2024).

Not only have the definitions and categories of race and ethnicity shifted dramatically over time, but the methods used to measure them have also evolved throughout the Census's history. From 1790 until 1950, enumerators administered the Census, recording race based on their own observations (Jensen and Jones, 2024). In 1960, the Census transitioned to a mail-based system, enabling respondents to self-report their race for the first time. Today, the principle of self-identification has become central to most official collections of race and ethnicity statistics.

## 2.2 Self-Reports vs. Administrative Records

Federal guidelines established by the Office of Management and Budget (OMB) designate self-reporting as the preferred method for collecting racial data (National Research Council, 2004). While self-reporting has become the standard in survey research, administrative records are often collected through different mechanisms. Recent research has documented significant inconsistencies between self-reported and administrative measures of race, which can result in underestimating racial disparities as well as underrepresented populations (Baron et al., 2024; Finlay et al., 2024; Jarrín et al., 2020).

What accounts for racial misclassification in administrative records? For one, administrative records often reflect a third party's perception of an individual's race rather than the individual's own identification – similar to the historical practice of Census enumerators recording race based on observation.[4] This perception can, in turn, depend on the

[4]There are other reasons as well: Finlay et al. (2024) documents how operational and bureaucratic complications impede correction of racial misreports in the justice system, while Luh (2023) finds evidence

observer's own identity. For example, Dobre and Witzen (2023) provide evidence that mortgage lenders are more likely to classify an applicant's race differently from their self-identification when the lender belongs to a different racial group. While third-party reports of race can vary across observers, self-reported race avoids this complication: only one individual's response matters.

## 2.3 Linking to Sidestep Imputation

Researchers may avoid imputation altogether if they can link datasets missing race with other datasets that contain self-reported race. For example, researchers have linked bank account records to voter file data (Ganong et al., 2023), credit scores to mortgage data (Albanesi and Vamossy, 2024), and tax data to ACS survey data (Choukhmane et al., 2023) to access self-reported race.

However, linkage is not always feasible because of administrative costs, data use restrictions, and privacy concerns. Laws protecting confidentiality can limit data sharing, leading some organizations to invest in custom internal imputation methods instead (King et al., 2023).[5]

## 2.4 Our Definitions

In this paper, we evaluate the performance of imputation methods against self-reported race. We focus on self-reported race because it is defined within a given individual and time, and it is observable in publicly available data (voter file data).

When we discuss race in this paper, we are combining race and ethnicity. We make this choice for two main reasons. First, beginning with the 2030 decennial Census, race and ethnicity will be measured through a single combined question, which the Census Bureau expects will improve data quality (U.S. Census Bureau, 2024a).[6] Second, for practical purposes, we are constrained by the design of existing imputation methods, all of which already combine race and ethnicity. We therefore follow both Census guidance and prevailing methodological practice in treating these concepts jointly.

that criminal justice actors (highway patrol officers) may intentionally misreport race to conceal racial bias.

[5] King et al. (2023) describe in detail the privacy-bias trade-off in the context of US federal agencies. See Appendix A.1.1 for examples of US government agencies using race imputation.

[6] See Appendix A.1.2 for more background on this Census change.

# 3 Voter File Data

To test the performance of different race imputation methods, we need a dataset that includes self-reported race as well as other variables that are commonly used for imputation (i.e., name and geography). To the best of our knowledge, the only public US data that include name, geography, and high rates of self-reported race are US voter files.[7]

Six US states collect data about voters' self-reported race: Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina. This practice stems from the 1965 Voting Rights Act, which prompted expanded data collection to counter state-level voter disenfranchisement (McCartan et al., 2025; Cruz and Hayes, 2009).

For our main results, we use a random sample of 500,000 observations from the 2023 North Carolina (NC) voter file.[8] We use the North Carolina file because it is free and easy to access, which makes our replication materials the most accessible to other researchers.[9]

# 4 Race Imputation Methods

We compare imputation methods applicable to publicly available US voter file data containing self-reported race. These methods use names, geographic information, or both – the variables most commonly available for imputation. We evaluate standard methods (surname-based imputation, geography-based imputation, and BISG) alongside recently developed methods (Zest Race Prediction, NamePrism, and BIRDiE).

All methods generate predicted probabilities that individuals belong to one of five racial groups (White, Black, Hispanic, Asian, and Other). For each method, we assign individuals to the racial group with the highest predicted probability.[10]

[7] There is one other public dataset that includes name, geography, and self-reported race: public PPP data for loans over $150,000. However, about 90% of the PPP loan applications are missing race information (Atkins et al., 2022), and there might have been strategic selection into reporting race at all (García and Darity Jr, 2022).

[8] One concern with this approach is that the patterns and results might differ if we used data from a different place. Indeed, prior research has shown imputation performance varies meaningfully by place (Decter-Frain, 2022). To address this concern, we use Appendix A.7 to show our results are similar if we instead use a random sample of 500,000 observations from the Florida voter file.

[9] For more background on the data, see Appendix A.2.

[10] An alternative approach uses probability thresholds to make assignments, which we explore in Appendix A.8.

## 4.1 Last name

Surname-based imputation uses Census surname frequency tabulations. We implement this using the 2000 and 2010 Census surname files (U.S. Census Bureau, 2022, 2021), which include surnames appearing at least 100 times (Kaplan, 2025).

## 4.2 Zip code

Geography-based imputation uses the racial composition of the area where individuals reside. We use zip codes of residence, with racial demographics from the 2019 5-year ACS. We use zip codes rather than finer geographic units to avoid geocoding complications.[11] Zip codes have population distributions similar to census tracts (Clark et al., 2022).

## 4.3 Bayesian Improved Surname Geocoding (BISG)

Bayesian Improved Surname Geocoding (BISG) imputes race based on both surname and geographic data. The BISG methodology is an application of naive Bayes prediction, where the key assumption is that surname and geolocation within racial groups are independent (Elliott et al., 2009).

The prior probability $p(i|j)$ is the proportion of people with surname $j$ who report as race $i$. This is updated with geographic data: $r(k|i)$ is the proportion of people reporting race $i$ who reside in zip code $k$. Then, $q(i|j,k)$ is the updated (posterior) probability of being of race $i$ given surname $j$ and zip code $k$:

$$q(i|j,k) = \frac{p(i|j)r(k|i)}{\sum_{s=1}^{6} p(s|j)r(k|s)}$$

Imai and Khanna (2016) improved BISG by incorporating voter registration data from six Southern states collecting self-reported race. We implement this updated version of BISG using zip codes and 2010 Census data (Curiel, 2024).

## 4.4 Zest Race Prediction (ZRP)

Zest Race Prediction (ZRP) uses machine learning techniques with first, middle, and last names plus geographic information (Matthews et al., 2024a). ZRP was trained on 2021

[11] Zip code is explicitly in the voter file data, while other geographic units of interest (county, tract, block) require geocoding addresses, which can introduce its own methods ambiguity. See Appendix A.3 for more information on geocoding complications.

voter files from three states (Florida, Georgia, and North Carolina) (Matthews et al., 2024a). We implement ZRP using zip code data (Matthews et al., 2024).

Early evidence suggests ZRP can improve on BISG in some dimensions. Matthews et al. (2024b) show higher classification accuracy in Alabama voter records and higher classification accuracy and more accurate counts for small business owners. Curiel and DeLuca (2024a) find ZRP performs especially well in racially diverse areas where BISG's Bayesian prior is less informative.

### 4.5 NamePrism

NamePrism uses machine learning to classify nationality and ethnicity from names (Ye et al., 2017). Unlike the other methods we compare, which are designed specifically for the US context, NamePrism was developed as an international classifier. It leverages 74 million labeled name-nationality pairs from email and Twitter data, plus 57 million contact lists, combining unsupervised learning with supervised classification. The method predicts one of 39 nationality classes, which are then mapped to six US race/ethnicity categories for US applications.

The developers show NamePrism outperforms competing international classifiers in individual-level classification (Ye et al., 2017). To our knowledge, it has not been systematically compared to US-specific methods like BISG. NamePrism has been growing in popularity and has been cited by over 150 papers as of August 2025.

### 4.6 BIRDiE

Unlike the methods above, BIRDiE (Bayesian Instrumental Regression for Disparity Estimation) is an adjustment method rather than direct imputation method (McCartan et al., 2025). BISG produces well-calibrated racial predictions, but residuals are often correlated with outcomes of interest, biasing racial disparity estimates. BIRDiE addresses this by using surnames as a high-dimensional instrumental variable for race.

Specifically, McCartan et al. (2025) assume the outcome is conditionally independent of surname given race, residence location, and other observed attributes. (This assumption is a type of exclusion restriction.) This implies that for two individuals in the same area, same racial group, and with the same observables, their surnames have no predictive power over the outcome.

We implement BIRDiE corrections using BISG as the baseline (McCartan, 2023). Since

BIRDiE aims to improve disparity measurement rather than individual-level prediction, we only evaluate its performance for racial gap estimation.

# 5 Dimensions of Race Imputation Performance

Race imputation methods can be evaluated along multiple dimensions. Canonical measures of performance are measures of individual-level classification: which methods avoid false negatives and false positives? Meanwhile, applied researchers typically impute race to answer one of two different questions: How large are racial disparities in economic outcomes? What is the racial composition of a population? These three goals – individual-level classification, disparity estimation, and demographic estimation – represent distinct dimensions of imputation performance, which we describe further below.

## 5.1 Individual-level Classification Performance

Individual-level classification performance is the primary focus in papers that introduce new race imputation methods or compare methods (Elliott et al., 2009; Decter-Frain, 2022; Curiel and DeLuca, 2024a; Chaturvedi and Chaturvedi, 2024; Chalavadi et al., 2025; Li, 2023). While applied researchers do not focus on individual-level prediction as their primary goal, this metric is frequently cited to justify methodological choices.

We evaluate individual-level performance using F1 scores, which are well-suited to multi-class classification problems like race imputation. F1 scores are a function of recall and precision, and can be computed for each racial group and aggregated for an overall performance measure. Below is how recall, precision, and F1 scores are defined for each specific racial group $g$:

1. Recall: The fraction of people of group $g$ who are imputed as group $g$. $Recall = \frac{TP}{TP+FN}$ where $TP$ is true positives and $FN$ is false negatives.
2. Precision: The fraction of people imputed as group $g$ who truly belong to group $g$. $Precision = \frac{TP}{TP+FP}$ where $TP$ is true positives and $FP$ is false positives.
3. F1 score: The harmonic mean of recall and precision: $F1 = \frac{2\times(precision\times recall)}{precision+recall}$.

These metrics are defined separately for each racial group. To summarize overall performance, we compute three types of aggregated F1 scores:

1. Micro average: Aggregates counts across all classes before calculating the metric: $F1^{micro} = \frac{TP}{TP+0.5\times(FP+FN)}$. This gives equal weight to each individual.

2. Macro average: Takes a simple average of each racial group's F1 score, giving equal weight to each group regardless of population size.
3. Weighted average: Weights each racial group's F1 score by its observed population in the sample.

## 5.2 Performance on Estimating Racial Gaps

The primary motivation for race imputation in economics is to estimate racial disparities, differences in economic outcomes between groups. Researchers use imputation for this purpose across many topics including taxes (Elzayn et al., 2025), credit access (Frame et al., 2021), and retirement savings (Coyne et al., 2024).

To evaluate performance on gap estimation, we require an outcome variable available in voter files. We use party registration, defining our outcome as registering as a Democrat. We begin by examining the Black-White gap, as studied by McCartan et al. (2025), since these are the two largest populations in the North Carolina voter file. We then extend the analysis to pairwise gaps between the four largest racial groups: Black-White, Hispanic-White, Asian-White, Black-Hispanic, Black-Asian, and Hispanic-Asian. This comprehensive approach reveals whether methods perform consistently across different racial gaps.

## 5.3 Performance on Estimating Racial Demographics

Researchers also use race imputation to describe a population's racial composition. Examples include describing the racial composition of evicted people (Hepburn et al., 2020), people killed by police Edwards et al. (2019), and people infected with COVID-19 (Labgold et al., 2021).

To evaluate demographic estimation performance, we compare the imputed percentage of each racial group with the observed percentages in the voter file. For a simple summary statistic, we calculate the Euclidean distance between the imputed and observed distributions (treating each as a point in 5-dimensional space). Smaller distances indicate better performance, with a distance of zero representing perfect accuracy.

# 6 Comparative Performance Results

We evaluate race imputation methods across three distinct performance dimensions: individual-level classification, racial gap estimation, and demographic estimation. If a race imputation method were 100% accurate (every imputed measure of race aligned with self-reported race), it would necessarily dominate on all three types of performance. However, when no imputation method is 100% accurate, we find the highest-performing method in one dimension may underperform in others. We therefore compare methods across different types of performance, discussing each dimension in Sections 6.1-6.3 before synthesizing findings in Section 6.4.

## 6.1 Individual-level Classification Performance

We begin by evaluating individual-level classification performance, the dimension most commonly emphasized in papers introducing new methods. For this dimension, we use F1 scores, which combine recall and precision and are well-suited to multi-class classification problems like race imputation. Figure 2a reports F1 scores by racial group, while Figure 2b reports global F1 scores aggregated using three different weighting schemes.

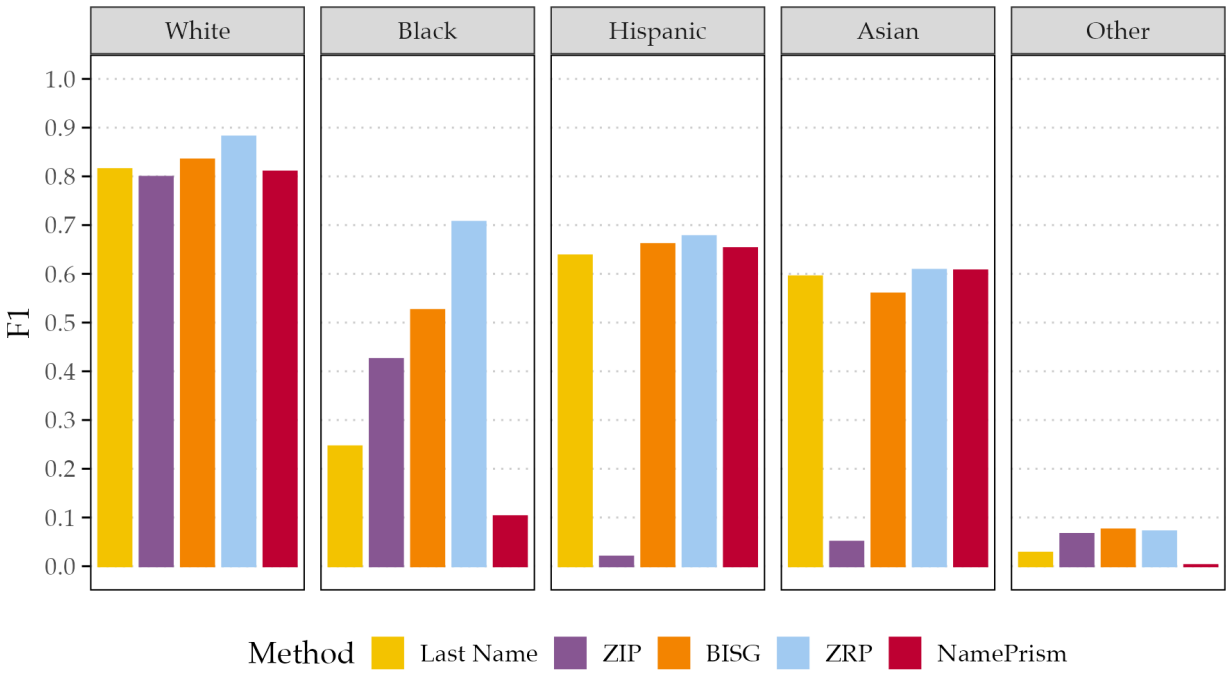
We compare both standard and newer methods for imputing race, as outlined in Section 4. The standard methods are surname imputation, Zip-based imputation (a geography-based approach), and BISG, represented by red, pink, and yellow bars, respectively. The newer methods are Zest Race Prediction (ZRP), shown in blue, and NamePrism, shown in green.

Figures 2a and 2b reveal several key patterns in the performance of standard imputation methods. First, methodological performance varies substantially by the racial group of interest. For example, Figure 2a shows that geography-based imputation outperforms surname imputation for the Black population, but yields near-zero F1 scores for the Hispanic and Asian populations. In contrast, surname imputation performs comparatively better for these groups. This pattern reproduces the established finding that geography-based methods are more effective for imputing Black identity, while surname-based methods are more effective for imputing Hispanic and Asian identity.[12]

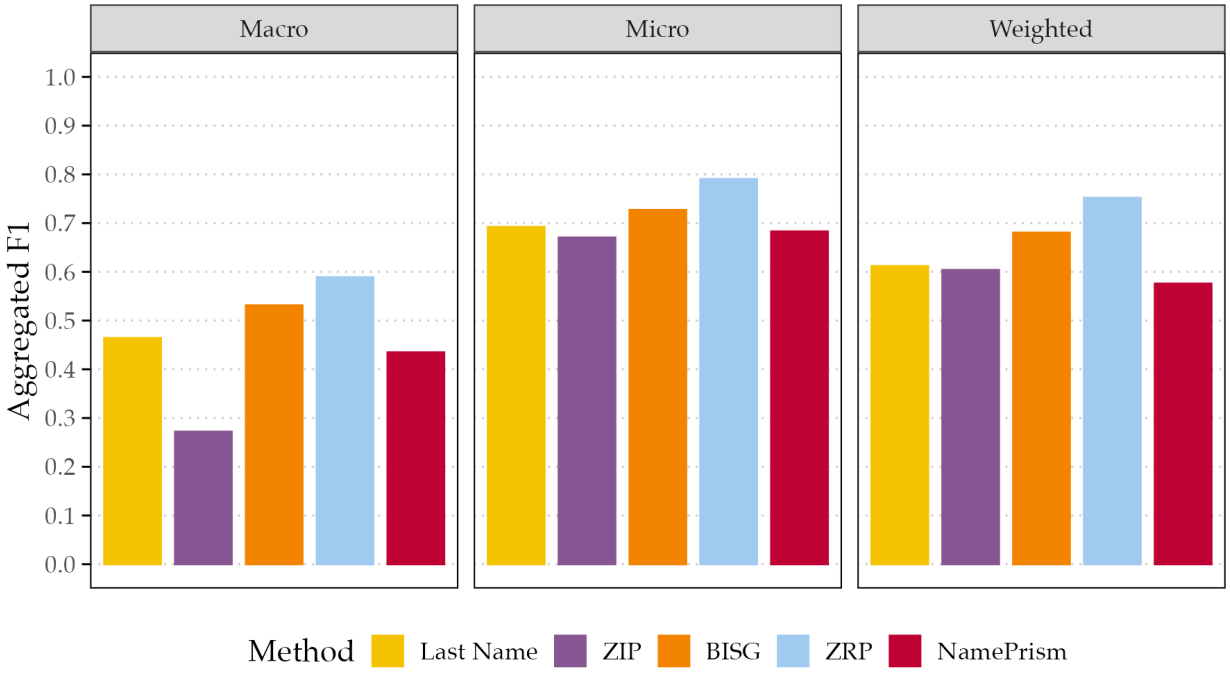
[12]This finding has been previously documented by Elliott et al. (2009) and Fiscella and Fremont (2006). This result is a consequence of the nature of geographic segregation and naming patterns in the US. Black-White segregation levels are higher than Asian-White or Hispanic-White segregation levels, which leads geography-based methods to work better for inferring Black identity (Iceland et al., 2010; Logan, 2013). Meanwhile, surname imputation performs well for inferring Asian identity due to distinctive naming patterns (Fiscella and Fremont, 2006; Comenetz, 2016).

Figure 2: F1 Score Performance Analysis by Race Imputation Method

(a) F1 Scores by Racial Group and Race Imputation Method




(b) Aggregated F1 Scores by Race Imputation Method




*Notes:* This figure shows F1 scores across racial groups and race imputation methods (Panel A) and aggregated F1 scores by method (Panel B) using a 500,000 random sample of the NC voter file.

Second, these figures demonstrate how BISG, which uses both surname and geography data, compares to using only surname or geography. In terms of aggregate measures, Figure 2b shows that BISG generates higher F1 statistics than surname or Zip methods regardless of weighting method. However, the aggregate picture masks important differences across underlying racial groups, shown in Figure 2a. BISG improves performance relative to the two simpler methods for most groups, but not all. The largest performance gains from using BISG appear for the Black population. Meanwhile, the gains to using BISG relative to surname are slight for the White and Hispanic populations, and BISG actually underperforms relative to the surname method for the Asian population. Therefore, the aggregate gains from BISG are driven by some racial groups more than others.

Finally, these figures extend beyond the standard methods by comparing two newer ML-based approaches to race imputation: ZRP and NamePrism. ZRP achieves the highest aggregated F1 score, regardless of weighting method (Figure 2b), and outperforms all other methods for White, Black, Hispanic, and Asian populations (Figure 2a). The gains over standard methods are especially notable for the Black population, where ZRP attains an F1 score of 0.7 compared to BISG's 0.5. In contrast, NamePrism often yields lower aggregated F1 scores than standard methods (Figure 2b), and its performance for the Black population is strikingly poor, with an F1 score of just 0.1 (Figure 2a).

## 6.2 Performance on Estimating Racial Gaps

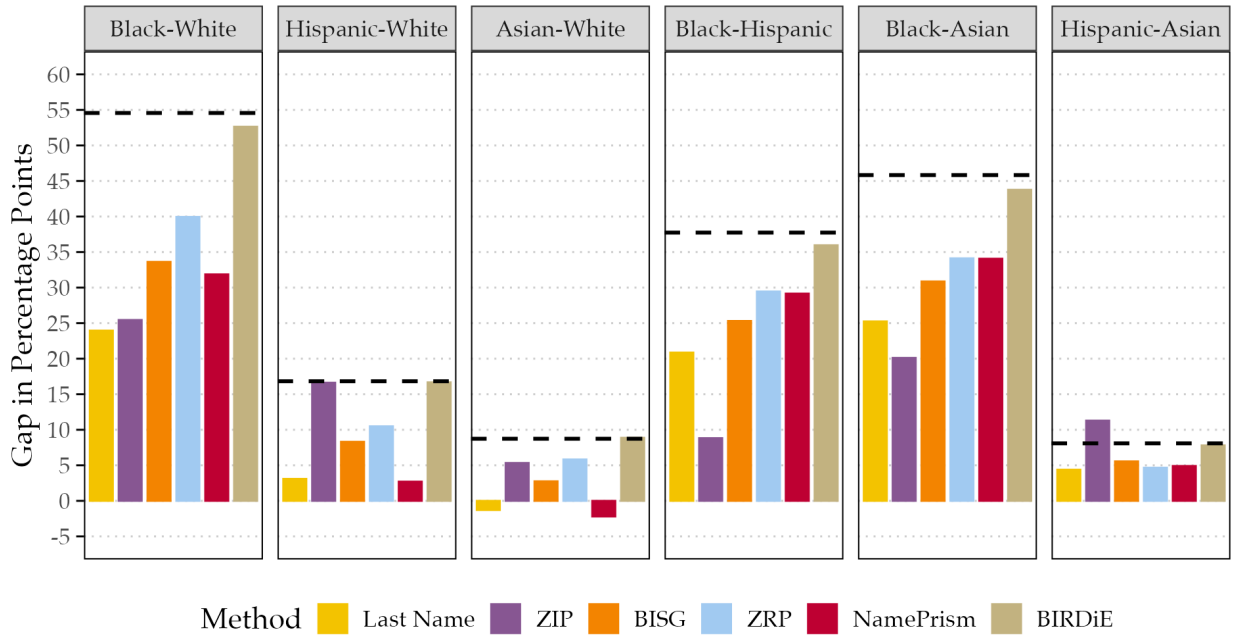
To evaluate performance on racial gap estimation, we focus on Democratic Party registration, one of our observable outcomes in the voter file. We begin by examining the Black-White gap since these are the two largest populations in the North Carolina voter file.

Figure 3 illustrates how estimates of different racial gaps vary across imputation methods. The gap between the largest two racial groups (Black and White) is shown on the left. In the observed data, 76% of Black registered voters are Democrats compared to 21% of White registered voters, yielding a 55-percentage point gap, which is demonstrated with the black dashed line. The standard methods – surname, Zip, and BISG – substantially underestimate this gap by 20 to 30 points. Among the newer direct imputation methods, NamePrism performs similarly to BISG, while ZRP produces a more accurate estimate (40 percentage points). Thus, ZRP comes closest to recovering the true gap among the direct imputation methods.

Importantly, these results demonstrate that methods can outperform in some dimensions

of performance even when they underperform in others. NamePrism outperforms surname imputation when estimating the Black-White gap (Figure 3), despite yielding lower aggregate F1 scores and lower White and Black F1 scores than surname imputation (Figures 2b and 2a). Therefore, higher F1 scores do not guarentee more accurate measures of racial gaps.

Figure 3: Racial Gaps in Democratic Party Registration by Race Imputation Method




*Notes:* This figure shows the estimated Black-White disparity in democratic party registration (Panel A) and all racial gaps in democratic party registration (Panel B) by different race imputation methods. The gray bar/dots represent the real gaps in the data.

The right-most beige bar in Figure 3 shows the estimated gap using BIRDiE adjustment.[13] Consistent with McCartan et al. (2025), we find that BIRDiE substantially improves estimates of the Black-White gap. Its estimate is within a few percentage points of the observed gap, which is an especially large improvement given that BISG underestimates the gap by about 20 points. While ZRP narrows the gap relative to BISG, its accuracy still falls well short of what the BIRDiE adjustment achieves.

Moving beyond the largest populations, we also examine performance across other racial gaps. Figure 3 demonstrates six estimated racial gaps (Black-White, Black-Hispanic, Black-Asian, Hispanic-White, Hispanic-Asian, and Asian-White) across all imputation methods, and contrasts these estimates with observed gaps (shown with black dashed horizontal

[13]Unlike the direct individual-level imputation methods discussed above, BIRDiE is not based on imputing race at the person level but instead applies an adjustment specifically for estimating group-level disparities (McCartan et al., 2025). For this reason, we only evaluate it in this performance section.

lines). Several insights emerge from these results. First, BIRDiE closely approximates the observed magnitudes for all racial gaps. Second, imputation can yield estimates in the wrong direction (e.g., NamePrism for the Hispanic-Asian gap and surname imputation for the Asian-White gap). Third, the relative performance of methods varies depending on the racial gap considered. For instance, Zip-based imputation produces the largest underestimate for the Black-Hispanic gap but performs best among direct imputation methods for the Hispanic-White gap. The ranking of methods is therefore not consistent across different racial gaps.

Overall, Figure 3 underscores the importance of imputation method choice for estimating racial disparities. Different methods can yield a wide range of estimates for a given gap, and can even reverse its direction. Moreover, the relative performance of methods varies across distinct racial gaps. Nevertheless, the results consistently highlight superior performance when using BIRDiE adjustment (McCartan et al., 2025).

### 6.3 Performance on Estimating Racial Demographics

Researchers also use race imputation to describe a population's racial composition. To evaluate demographic estimation performance, we compare the imputed percentage of each racial group with the observed percentages in the voter file. For a simple summary statistic, we report the Euclidean distance between the imputed and observed distributions (treating each as a point in 5-dimensional space).

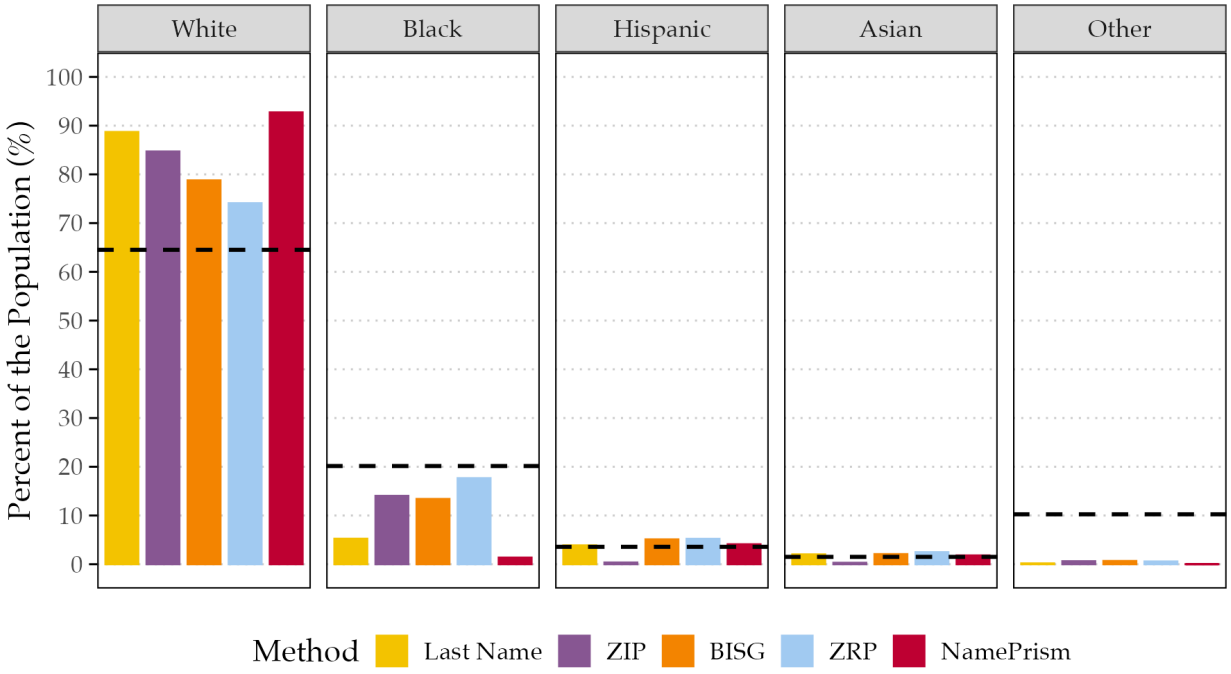
Figure 4a demonstrates several consistent patterns across all methods. First, all methods overestimate the percent of White people in the population, and underestimate the percent of Black people in the population. The most extreme case is NamePrism, which estimates 92% of the population is White and only 1% is Black, while the observed rates are 65% and 20%. The closest method to observed rates is ZRP, which yields results much closer to observed rates but still overestimates the White population (by 10 percentage points) and underestimates the Black population (by 2 percentage points).

Second, all methods perform poorly at estimating the percent of the population in the Other category. All methods estimate 0-1% of the population belongs to an Other racial group, when the observed rate is 10%. This poor performance reflects the fact that this group is not well-defined and is treated as the residual of the main four racial groups in imputation methods.

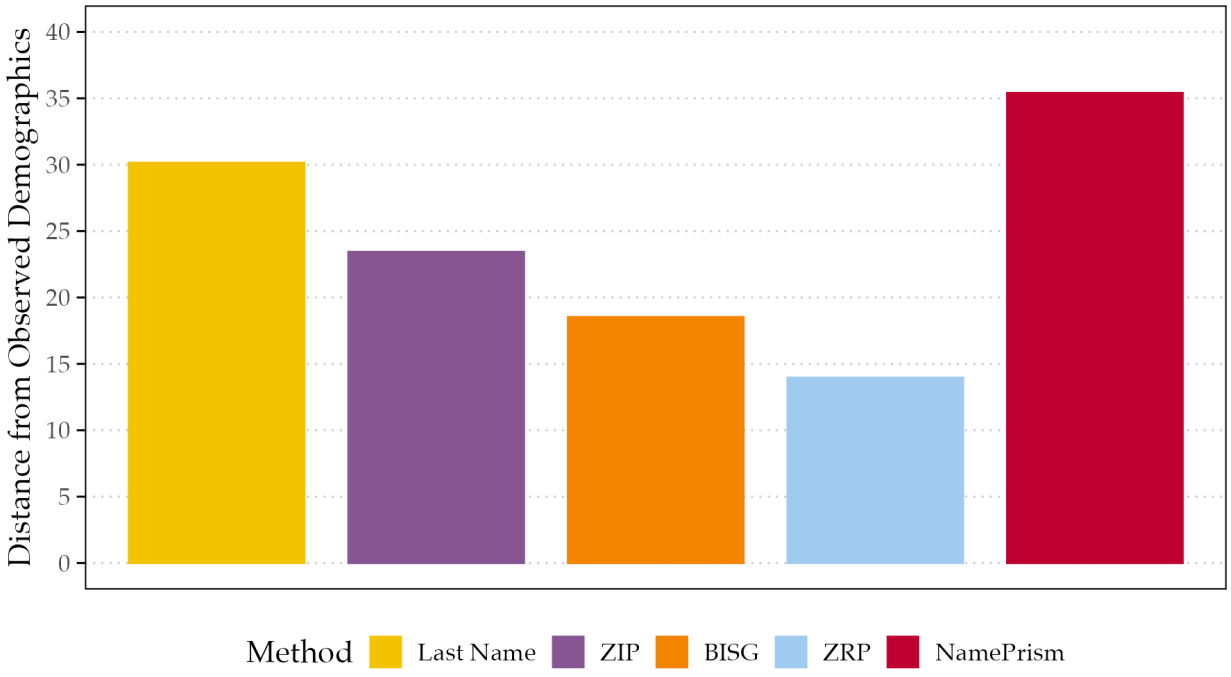
Using the Euclidean distance summary statistic, Figure 4b shows that ZRP comes closest to the observed population rates (with a distance of 13.96), followed by BISG (with a distance

Figure 4: Racial Demographics Performance by Race Imputation Method

(a) Simulated Racial Demographics by Race Imputation Method




(b) Distance from Observed Demographics by Race Imputation Method




*Notes:* Panel (a) shows the simulated racial demographics by race imputation method. The black horizontal dashed lines are the observed rates for each racial group. Panel (b) shows the euclidean distance (in 5-space) between the simulated demographic rates and the observed rates in the data.

of 18.53). Meanwhile, the ML-based method NamePrism is furthest from the observed rates (with a distance of 35.41), performing worse than simple surname-based imputation.

Once again, these results demonstrate that race imputation methods can outperform in some dimensions while underperforming in others. NamePrism performs worse than surname imputation in describing the population racial composition (Figure 4), but performs better when estimating the Black-White racial gap (Figure 3).

## 6.4 Cross-Dimensional Performance Patterns

We compare race imputation methods across three performance dimensions: individual-level classification, racial gap estimation, and demographic estimation. Standard methods (surname-based, geography-based, and BISG) often perform poorly on dimensions most relevant for applied research. All methods systematically overcount White people relative to other groups. Moreover, many methods substantially underestimate the magnitude of racial gaps, and some even reverse the direction of these gaps.

However, our results also demonstrate that two new methods make notable progress on the shortcomings of standard methods. First, Zest Race Prediction (ZRP) (Matthews et al., 2024) outperforms other direct imputation methods on individual-level classification (Figure 2b) and describing demographics (Figure 4). ZRP also outperforms other direct imputation methods for some racial gaps, but not all (Figure 3). The second new method – BIRDiE (McCartan et al., 2025) – is a correction method specifically designed for estimating racial disparities. We find that BIRDiE yields very accurate measures of racial disparities, and does so consistently across different racial group combinations (Figure 3).

We also uncover important shortcomings with new methods that are gaining traction in applied research. The NamePrism method (Ye et al., 2017), which has been cited by over 150 papers (as of August 2025), does not consistently outperform basic surname-based imputation despite its use of proprietary data and more sophisticated techniques. It yields lower F1 scores (Figure 2b) and less accurate population estimates (Figure 4) than surname-based imputation. We find evidence that this poor performance is overwhelmingly driven by its inability to accurately identify Black individuals (Figure 2a). This shortcoming likely reflects NamePrism's origins: it was developed as an international nationality classifier and adapted for US race classification, but this adaptation appears to fall short for Black people in the US. Therefore, newer methods leveraging machine learning are not always an improvement over existing methods, and methods designed for international contexts may not translate effectively to US racial classification.

Our results reveal a key methodological insight: methods can outperform on one dimension while underperforming on others. For example, NamePrism outperforms surname imputation when estimating the Black-White gap, despite yielding lower F1 scores for both White and Black populations and lower aggregate F1 scores. Similarly, ZRP dominates on individual classification but doesn't consistently outperform on all racial gaps. This pattern has important implications for empirical practice. High individual-level classification accuracy (measured by F1 scores) does not guarantee superior racial gap or descriptive statistics. Researchers selecting methods based on F1 scores alone, a common empirical practice, risk poor performance on the dimension that matters for their analysis. These findings highlight the need for broader study of imputation methods to document their assumptions, limitations, and performance across multiple dimensions.[14]

# 7 Conclusion

Race imputation has become common when studying economic inequality, with citations to standard methods more than tripling between 2020 and 2024. Researchers use imputation across topics including taxes, housing, credit, and retirement. Despite this prevalence and the variety of available methods, systematic comparisons remain rare. This paper addresses that gap by evaluating both standard and newly developed methods across three distinct performance dimensions: individual-level classification, racial gap estimation, and demographic estimation. Our key finding is that no single method dominates across all dimensions – methods can excel on one dimension while underperforming on others.

For practitioners, our results provide clear guidance. Researchers estimating racial disparities should use BIRDiE adjustment, which consistently produces accurate gap estimates across all racial group comparisons we examine. For demographic estimation, ZRP shows promise as a newer ML-based method. However, not all ML-based methods are equal: NamePrism does not consistently outperform basic surname-based imputation in US contexts and performs particularly poorly for Black populations. More fundamentally, our results challenge the common practice of selecting methods based solely on individual-level classification metrics. We demonstrate that superior F1 scores do not guarantee superior racial gap or demographic estimates.

Race imputation is just one example of how data-driven predictions are underlying a

[14] We discuss and compare additional race imputation methods in a number of Appendices: Appendix A.4.1 compares ZRP to other recent ML-based methods, Appendix A.5 discusses methods that require more input variables, and Appendix A.6 discusses proprietary race imputation methods.

growing amount of empirical research. Studying the specifics of these methods is crucial, as the proliferation of individual-level data rapidly expands our capacity for imputation.

# References


Albanesi, S. and D. F. Vamossy (2024). Credit scores: Performance and equity. Technical report, National Bureau of Economic Research.

Argyle, B., S. Indarte, B. Iverson, and C. Palmer (2025). Racial disparities and bias in consumer bankruptcy. Technical report, National Bureau of Economic Research.

Argyle, L. P. and M. Barber (2022). Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records. *American Political Science Review*, 1–9.

Atkins, R., L. Cook, and R. Seamans (2022). Discrimination in lending? evidence from the paycheck protection program. *Small Business Economics 58*(2), 843–865.

Barber, M. and L. P. Argyle (2023). Replication Data for: Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records.

Baron, E. J., J. J. Doyle Jr, N. Emanuel, P. Hull, and J. Ryan (2024). Unwarranted Disparity in High-Stakes Decisions: Race Measurement and Policy Responses. *NBER Chapters*.

Berry, A., E. M. Maloney, and D. Neumark (2024). The Missing Link? Using LinkedIn Data to Measure Race, Ethnic, and Gender Differences in Employment Outcomes at Individual Companies. Technical report, National Bureau of Economic Research.

Bleemer, Z. (2025). The vietnam war and racial integration.

Box-Couillard, S. and P. Christensen (2024). Racial housing price differentials and neighborhood segregation. Technical report, National Bureau of Economic Research.

Chalavadi, S., A. Pastor, and T. Leitch (2025). Lstm+ geo with xgboost filtering: A novel approach for race and ethnicity imputation with reduced bias. *arXiv preprint arXiv:2504.21259*.

Chaturvedi, R. and S. Chaturvedi (2024). It's All in the Name: A Character-Based Approach to Infer Religion. *Political Analysis 32*(1), 34–49.

Chintalapati, R., S. Laohaprapanon, and G. Sood (2018). Predicting Race and Ethnicity From the Sequence of Characters in a Name.

Choukhmane, T., J. Colmenares, C. O'Dea, J. Rothbaum, and L. Schmidt (2023). Who Benefits from Retirement Saving Incentives in the US? Evidence on Racial Gaps in Retirement Wealth Accumulation.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–980.

Chyn, E. and K. Haggag (2023). Moved to vote: The long-run effects of neighborhoods on political participation. *Review of Economics and Statistics 105*(6), 1596–1605.

Clark, J. T., J. A. Curiel, and T. S. Steelman (2021). New levels of racial estimation.

Clark, J. T., J. A. Curiel, and T. S. Steelman (2022). Minmaxing Of Bayesian Improved Surname Geocoding And Geography Level Ups In Predicting Race. *Political Analysis 30*(3), 456–462.

Cohle, Z. and A. Ortega (2023). The Effect of the Opioid Crisis on Patenting. *Journal of Economic Behavior & Organization 214*, 493–521.

Cohn, D. (2010). Race and the census: The "negro" controversy. *Pew Research Center*.

Comenetz, J. (2016). Frequently occurring surnames in the 2010 census. *United States Census Bureau*, 1–8.

Consumer Financial Protection Bureau (2014). Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment. *Consumer Financial Protection Bureau Working Paper*.

Cook, L. D., M. Marx, and E. Yimfor (2025). Funding black high-growth startups.

Costello, R., P. DeFilippes, R. Fisher, B. Klemens, and E. Y. Lin (2024). Marriage penalties and bonuses by race and ethnicity: An application of race and ethnicity imputation. In *AEA Papers and Proceedings*, Volume 114, pp. 644–648. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Coyne, D., I. Fadlon, and T. Porzio (2024). Measuring Valuation of Liquidity with Penalized Withdrawals.

Cronin, J.-A., P. DeFilippes, and R. Fisher (2023). *Tax expenditures by race and hispanic ethnicity: An application of the us treasury department's race and hispanic ethnicity imputation*. US Department of the Treasury, Office of Tax Analysis.

Cruz, J. and J. Hayes (2009). Adding Race and Ethnicity: Electoral Data Collection Practice and Prospects for New York State.

Curiel, J. A. and K. DeLuca (2024a). The power of characters: Evaluating machine learning-modified bayesian improved surname geocoding inference of race in redistricting. *State Politics & Policy Quarterly 24*(3), 300–321.

Curiel, J. A. and K. DeLuca (2024b). The power of characters: Evaluating machine learning-modified bayesian improved surname geocoding inference of race in redistricting. *State Politics & Policy Quarterly*, 1–22.

Decter-Frain, A. (2022). How Should We Proxy for Race/Ethnicity? Comparing Bayesian Improved Surname Geocoding to Machine Learning Methods. *arXiv preprint arXiv:2206.14583*.

Diamond, R. and W. F. Diamond (2024). Racial differences in the total rate of return on owner-occupied housing. Technical report, National Bureau of Economic Research.

Dobre, A. and B. H. Witzen (2023). Mortgage Lender Misclassification of Applicant Race and Ethnicity: Evidence from HMDA Data. *Consumer Financial Protection Bureau Office of Research Working Paper* (23-04).

Edwards, F., H. Lee, and M. Esposito (2019). Risk of Being Killed by Police Use of Force in the United States by Age, Race-Ethnicity, and Sex. *Proceedings of the national academy of sciences 116*(34), 16793–16798.

Einav, L. and J. Levin (2014). Economics in the age of big data. *Science 346*(6210), 715–721.

Elliott, M. N., P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie (2009). Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities. *Health Services and Outcomes Research Methodology 9*, 69–83.

Elzayn, H., E. Smith, T. Hertz, C. Guage, A. Ramesh, R. Fisher, D. E. Ho, and J. Goldin (2025). Measuring and mitigating racial disparities in tax audits. *The Quarterly Journal of Economics 140*(1), 113–163.

Feigenberg, B. and C. Miller (2022). Would eliminating racial disparities in motor vehicle searches have efficiency costs? *The Quarterly Journal of Economics 137*(1), 49–113.

Finlay, K., E. Luh, and M. G. Mueller-Smith (2024, July). Race and Ethnicity (Mis)measurement in the U.S. Criminal Justice System.

Fiscella, K. and A. M. Fremont (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research 41*(4p1), 1482–1500.

Fisher, R. (2023). Estimation of Race and Ethnicity by Re-Weighting Tax Data. *The Department of Treasury Office of Tax Analysis. Technical Working Paper 11*.

Frake, J., R. Hurst, and M. Kagan (2024). Partisan segregation in the us workplace is large and rising. *Available at SSRN 4639165*.

Frame, W. S., R. Huang, E. J. Mayer, and A. Sunderam (2021). The Impact of Minority Representation at Mortgage Lenders. *SMU Cox School of Business Research Paper* (22-08).

Fremont, A., J. S. Weissman, E. Hoch, and M. N. Elliott (2016). When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities. *RAND Health Quarterly 6*(1).

Ganong, P., D. Jones, P. Noel, D. Farrell, F. Greig, and C. Wheat (2023). Wealth, Race, and Consumption Smoothing of Typical Income Shocks.

García, R. E. and W. A. Darity Jr (2022). Self-reporting race in small business loans: A game-theoretic analysis of evidence from ppp loans in durham, nc. In *AEA Papers and Proceedings*, Volume 112, pp. 299–302. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Goldsmith-Pinkham, P. (2024). Tracking the credibility revolution across fields.

Goncalves, F. and S. Mello (2021). A few bad apples? racial bias in policing. *American Economic Review 111*(5), 1406–1441.

Greengard, P. and A. Gelman (2024). An Improved BISG for inferring Race from Surname and Geolocation. *arXiv preprint arXiv:2304.09126*.

Greenwald, D. L., S. T. Howell, C. Li, and E. Yimfor (2024). Regulatory arbitrage or random errors? implications of race prediction algorithms in fair lending analysis. *Journal of Financial Economics 157*, 103857.

Harris, A. (2020). Using Bayesian Improved Surname Geocoding (BISG) to Classify Race and Ethnicity in Administrative Employment Data by Industry: A Validation Study. https://ww2.amstat.org/meetings/jsm/2020/onlineprogram/AbstractDetails.cfm?abstractid=311006. Accessed on: April 19, 2024.

Hepburn, P., R. Louis, and M. Desmond (2020). Racial and Gender Disparities Among Evicted Americans. *Sociological Science 7*, 649–662.

Howell, S. T., T. Kuchler, D. Snitkof, J. Stroebel, and J. Wong (2024). Lender Automation and Racial Disparities in Credit Access. *The Journal of Finance 79*(2), 1457–1512.

Hurtado, A. and J. Sakong (2023). The effect of minority bank ownership on minority credit. *George J. Stigler Center for the Study of the Economy & the State Working Paper* (325).

Iceland, J., K. A. Goyette, K. A. Nelson, and C. Chan (2010). Racial and ethnic residential

segregation and household structure: A research note. *Social Science Research 39*(1), 39–47.

Imai, K. and K. Khanna (2016). Improving Ecological Inference By Predicting Individual Ethnicity From Voter Registration Records. *Political Analysis 24*(2), 263–272.

Imai, K., S. Olivella, and E. T. Rosenman (2022). Addressing Census Data Problems In Race Imputation Via Fully Bayesian Improved Surname Geocoding And Name Supplements. *Science Advances 8*(49), eadc9824.

Jarrín, O. F., A. N. Nyandege, I. B. Grafova, X. Dong, and H. Lin (2020). Validity of Race and Ethnicity Codes in Medicare Administrative Data Compared with Gold-Standard Self-Reported Race Collected During Routine Home Health Care Visits. *Medical care 58*(1), e1–e8.

Jensen, E. and N. Jones (2024). Measuring the racial and ethnic composition and diversity of the united states population: Historical challenges and contemporary opportunities. Technical report, National Bureau of Economic Research, Inc.

King, G. (1995). Replication, replication. *PS: Political Science & Politics 28*(3), 444–452.

King, J., D. Ho, A. Gupta, V. Wu, and H. Webley-Brown (2023). The privacy-bias tradeoff: Data minimization and racial disparity assessments in us government. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 492–505.

Labgold, K., S. Hamid, S. Shah, N. R. Gandhi, A. Chamberlain, F. Khan, S. Khan, S. Smith, S. Williams, T. L. Lash, et al. (2021). Estimating the unknown: greater racial and ethnic disparities in covid-19 burden after accounting for missing race and ethnicity data. *Epidemiology 32*(2), 157–161.

Letzler, R., R. Sandler, A. Jaroszewicz, I. Knowles, and L. M. Olson (2017). Knowing when to quit: Default choices, demographics and fraud. *The Economic Journal 127*(607), 2617–2640.

Li, C. (2023). Can we trust race prediction? *arXiv preprint arXiv:2307.08496*.

Loewenstein, M., D. Piccone, and A. Polivka (2024). Measuring potential effects of the proposed race and ethnicity questions in the current population survey. Technical report, National Bureau of Economic Research.

Logan, J. R. (2013). The persistence of segregation in the 21st century metropolis.

Luh, E. (2023). Not So Black and White: Uncovering Racial Bias from Systematically Misreported Trooper Reports. *Available at SSRN 3357063*.

Matthew, A., A. R. Miller, and C. Tucker (2024, May). Algorithmic Bias and Historical Injustice: Race and Digital Profiling.

Matthews, K., P. Zak, A. Li, C. Williams, S. Kamkar, and J. Budzik (2024a). ZRP Model Development Documentation. https://github.com/zestai/zrp/blob/main/model_report.rst.

Matthews, K., P. Zak, A. Li, C. Williams, S. Kamkar, and J. Budzik (2024b). ZRP Validation Experiments' Results. https://github.com/zestai/zrp/blob/main/supporting_docs/validation_experiment_results.rst.

McCartan, C., R. Fisher, J. Goldin, D. E. Ho, and K. Imai (2025). Estimating Racial Disparities When Race Is Not Observed. *Journal of the American Statistical Association* (Forthcoming), 1–22.

McCartan, C., J. Goldin, D. E. Ho, and K. Imai (2023). Estimating Racial Disparities When Race Is Not Observed. *arXiv preprint arXiv:2303.02580*.

McKinney, K. L. and J. M. Abowd (2024). Estimating the potential impact of combined race and ethnicity reporting on long-term earnings statistics. Technical report, National Bureau of Economic Research.

Medina, J. and R. Lai (2023). An american puzzle: Fitting race in a box. *The New York Times*.

NamSor (2025a). Explore our plans. Accessed: 2025-02-04.

NamSor (2025b). Find out the origin of a name. Accessed: 2025-02-04.

NamSor (2025c). Namsor, name checker for gender, origin and ethnicity determination. Accessed: 2025-02-04.

National Research Council (2004). Defining race. In *Measuring Racial Discrimination*, Chapter 2, pp. 25–38. Washington, DC: The National Academies Press.

Parasurama, P. (2021). raceBERT – A Transformer-based Model for Predicting Race and Ethnicity from Names. *arXiv preprint arXiv:2112.03807*.

Rosenman, E. T., S. Olivella, and K. Imai (2023). Race and ethnicity data for first, middle, and surnames. *Scientific Data 10*(1), 299.

Sebo, P. (2023). How well does namsor perform in predicting the country of origin and ethnicity of individuals based on their first and last names? *Plos one 18*(11), e0294562.

Sood, G. (2022). Florida Voter Registration Data (2017 and 2022). https://doi.org/10.7910/DVN/UBIG3F.

Sorbero, M. E., R. Euller, A. Kofner, and M. N. Elliott (2022). *Imputation of Race and Ethnicity in Health Insurance Marketplace Enrollment Data, 2015-2022 Open Enrollment Periods*. RAND.

Starr, P. and C. Pao (2024). The multiracial complication: The 2020 census and the fictitious multiracial boom. *Sociological Science 11*, 1107–1123.

U.S. Census Bureau (2021). Frequently occurring surnames from the census 2010. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html. Accessed: 2025-08-13.

U.S. Census Bureau (2022). Frequently occurring surnames from the census 2000. https://www.census.gov/topics/population/genealogy/data/2000_surnames.html. Accessed: 2025-08-13.

U.S. Census Bureau (2024a). Updates to race/ethnicity standards for our nation. Accessed: May 21, 2025.

U.S. Census Bureau (2024b). What updates to omb's race/ethnicity standards mean for the census bureau. Accessed: May 21, 2025.

Ventura, I. M. and R. D. Flores (2025). The "rise" of multiracials? examining the growth in multiracial identification in the 2020 us census. *RSF: The Russell Sage Foundation Journal of the Social Sciences 11*(1), 44–64.

Whitsel, E. A., P. M. Quibrera, R. L. Smith, D. J. Catellier, D. Liao, A. C. Henley, and G. Heiss (2006). Accuracy of commercial geocoding: assessment and implications. *Epidemiologic Perspectives & Innovations 3*(1), 8.

Ye, J., S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, and S. Skiena (2017). Nationality Classification Using Name Embeddings. *arXiv*.

## Software References


Curiel, J. (2024). zipWRUext: Package for the wru ZIP code extension. Accessed: 2025-09-09.

Kaplan, J. (2025). predictrace: Predict the Race and Gender of a Given Name Using Census and Social Security Administration Data. Version 2.0.1. Available online at https://jacobkap.github.io/predictrace/.

Matthews, K., P. Zak, A. Li, C. Williams, S. Kamkar, and J. Budzik (2024). Zest Race Predictor. https://github.com/zestai/zrp.

McCartan, C. (2023). BIRDiE: Bayesian Instrumental Regression for Disparity Estimation.

Skiena, S. (2025). NamePrism API. https://www.name-prism.com/api. Accessed: 2025-10-10.

United States Census Bureau (2025, September). Census Geocoder. https://geocoding.geo.census.gov/geocoder/. Accessed: 2025-09-16.

# Supplemental Appendix

## A.1 Further Background

### A.1.1 Race Imputation in US Government Agencies

Race imputation has been used across multiple US government agencies beyond the US Treasury: the Consumer Financial Protection Bureau (CFPB) to assess fair lending (Consumer Financial Protection Bureau, 2014), the Equal Employment Opportunity Commission (EEOC) to analyze demographic disparities in labor discrimination cases (Harris, 2020), and the Department of Health & Human Services (HHS) and the Center for Medicare and Medicaid Services (CMS) to study healthcare inequities (Sorbero et al., 2022; Fremont et al., 2016). The US Court of Appeals has also upheld the use of race imputation in federal voting rights cases (Greengard and Gelman, 2024).

### A.1.2 Combined Race and Ethnicity in the US Census

The Census reports that growth in the "Some Other Race" category has largely been driven by Hispanic respondents who do not see their identity reflected in the existing race options (U.S. Census Bureau, 2024b). A combined race and ethnicity question is expected to reduce reliance on this residual category by encouraging respondents to select from the available choices. This change is also planned for earlier adoption in the American Community Survey, beginning in 2027 (U.S. Census Bureau, 2024a). For analyses of how this shift may affect measurement, see Loewenstein et al. (2024) for impacts on the Current Population Survey (CPS) and McKinney and Abowd (2024) for implications for long-term earnings statistics.

## A.2 Data Details

Collecting self-reported race was mandatory in Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina for many years but became optional after the Supreme Court's Shelby County vs. Holder decision. Even though the collection of self-reported race data is now optional, race is still self-reported 92.5% of the time in our North Carolina random sample.

We use a 500,000 random sample instead of the whole voter file in order to avoid hold-up times with more computationally difficult methods. In particular, we can impute race

quickly using all methods in the main text except for one: NamePrism. NamePrism is available as a free web service for researchers (Skiena, 2025), but the API limits the request rate to 60 requests per minute. We limit the sample to a random 500,000 observation subset of the 8 million observations in the North Carolina data, so that we can test NamePrism against the other methods.

The race categories in the NC voter file do not align with standard race categories yielded by race imputation methods. The racial groups in the NC voter file are: Asian, Black or African American, American Indian or Alaska Native, Two or more races, Other, Native Hawaiian or Pacific Islander, Undesignated, White, and Hispanic. Meanwhile, the standard racial categories across the imputation methods we study are: White, Asian, Black, Hispanic, and Other. Therefore, to map the race categories in the voter file data to standard categories for imputation, we follow Imai and Khanna (2016) and classify Native Hawaiian or Pacific Islander people as Asian, and people who report American Indian or Alaska Native, Two or more races, Other, and Undesignated as Other.

## A.3 Geocoding Complications

Geocoding addresses can introduce a number of complications and issues. First, geocoding leads to missing data because not all addresses will be geocoded. In our context, we lose 15% of observations when we batch geocode using the Census API (because of missing address or failure of geocoding from address to tract or block).

Second, geocoding generates its own methods ambiguity. Missing rates and accuracy varies by geocoding methods (Clark et al., 2021; Whitsel et al., 2006). Examples of prioprietary geocoders are Google, HERE, and ESRI. Meanwhile, free geocoding can be implemented by using the Census API, which bases geocoding on Census Bureau TIGER/Line files (United States Census Bureau, 2025). In fact, some papers that study race imputation with voter file data use L2 proprietary data because it already includes geocodings (McCartan et al., 2023; Rosenman et al., 2023).

Third, geocoding can be financially (if using a proprietary vendor) or computationally expensive (if using open source methods) (Clark et al., 2021). Therefore, to use as much data as possible, avoid methodological ambiguity, and avoid financial/computational costs, we avoid geocoding and use zip code for all the imputation methods that use geography.

## A.4 Comparing ZRP to other ML Methods

This appendix compares ZRP to two alternative machine learning algorithms that can impute race in the US. We evaluate their performance using voter registration data from North Carolina and Florida. We use both datasets because one method (CNN) was trained on North Carolina data, while the other (RaceBERT) was trained on Florida data.

We find that in most cases (across performance metrics and datasets), ZRP dominates both CNN and RaceBERT, though there are a few exceptions. Moreover, as expected, CNN outperforms RaceBERT in North Carolina and RaceBERT outperforms CNN in Florida. This is not surprising because CNN was trained on North Carolina data, while RaceBERT was trained on Florida data.

In the main text of the paper, we only present ZRP. We do this for two reasons. First, the results in Section A.4.2 below show that ZRP dominates CNN and RaceBERT on most metrics of performance (with a few exceptions). Second, ZRP is more attractive from an intuitive perspective because ZRP is not trained on a single state, meanwhile CNN was only trained on North Carolina and RaceBERT was only trained on Florida. Interested readers can consult this appendix for the detailed results across the three methods and two different datasets, but the main text only reports ZRP for succinctness.

### A.4.1 Two ML-Based Alternatives to ZRP

**RaceBERT**: Parasurama (2021) introduces a transformer-based model to impute race, which they call RaceBERT.[15] The model is an iteration of a popular type of transformer-based model, called a BERT (Bidirectional Encoder Representations from Transformers), which specializes in understanding the context of words. The model is trained on voter registration data from the state of Florida.[16]

The authors explain that they achieve state-of-the-art results in race prediction with high classification success. RaceBERT is very similar to a previous approach to race imputation by Chintalapati et al. (2018), called ethnicolr.[17] This previous approach used a Long Short Term Memory (LSTM) classifier instead of a transformer-based model. LSTM classifiers are recurrent neural networks built to handle data sequences and remember information over extended periods. The authors explain the RaceBERT method outperforms ethnicolr in terms of individual classification performance.

[15] RaceBERT is available as a python package online here: https://github.com/parasurama/raceBERT
[16] The authors used the February 2017 version of the FL voter file.
[17] Ethnicolr is available here: http://github.com/appeler/ethnicolr

**CNN**: CNN uses character-sequence-based machine learning to infer religion or race/ethnicity from names (Chaturvedi and Chaturvedi, 2024). The character patterns learned by the models work well for detecting identity groups for whom names may have different linguistic roots. The preferred model from Chaturvedi and Chaturvedi (2024) is a character-based convolutional neural network (CNN).[18] The method was initially developed to infer religion in India but the paper's online appendix extends their approach to infer race in US data using North Carolina voter file data.

When geography data is available, the character-based models can be incorporated within the existing BISG packages. (Chaturvedi and Chaturvedi, 2024) provide evidence that their CNN method with geography improves performance over the BISG baseline in terms of individual-level performance and demographic estimation. The CNN method was trained with the North Carolina voter file.[19] We employ the CNN method that uses geography data in this appendix.

**ZRP**: ZRP employs a machine-learning model (Extreme Gradient Boosting, or XGBoost) to estimate race using a combination of first, middle, and last names along with granular location data from the US Census. It leverages address geocoding to identify Census block groups, tracts, or ZIP codes using the Census ARCGIS TIGER/Line Shapefiles, pairing this information with updated demographic data from the American Community Survey (ACS). Unlike RaceBERT and CNN, ZRP is not trained on data from a single state. Instead, it is trained on 3 states: North Carolina, Georgia, and Florida (the 2021 voter registration files). ZRP also benefits from more recent ACS data compared to standard BISG methods that rely on the 2010 Census.

### A.4.2 Results Discussion

We report results across these three methods across our main three types of performance: individual-level performance, racial gaps, and population demographics. In terms of datasets, we compare the results using North Carolina and Florida voter files. For North Carolina, we use the same 500,000 random sample as we do in the main text. In addition, we also use a 500,000 random sample from the Florida voter file.[20]

Figure A.1 compares F1 scores by racial group across the three machine learning race

[18] They train language model, logistic regression, SVM, and CNN classifiers. They show that CNN classifiers have the best performance in terms of individual classification and estimating demographics, outperforming the standard BISG methods.

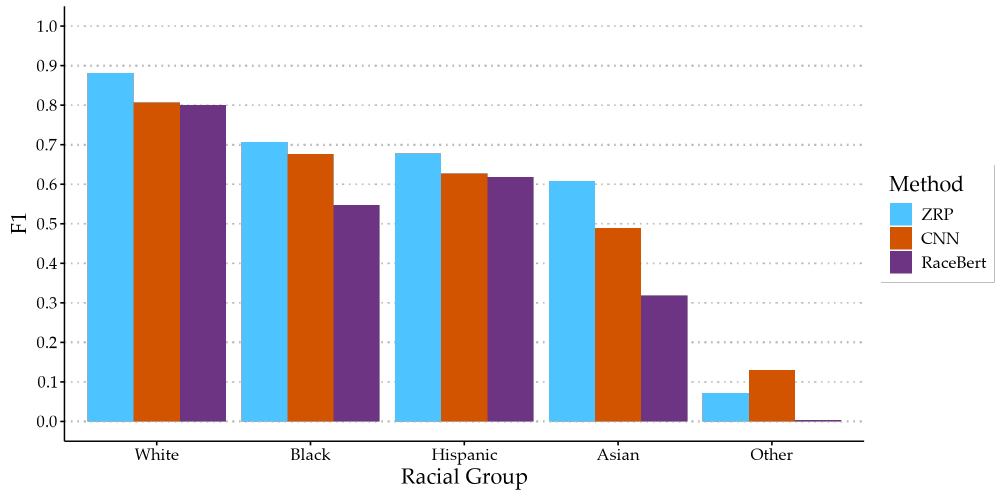
[19] The paper does not specify which year of the NC voter file the authors used.

[20] Both CNN and RaceBERT have zero missing rates for the North Carolina and Florida datasets. For ZRP the missing race in Florida is zero while the missing rate in North Carolina is only 0.01%.

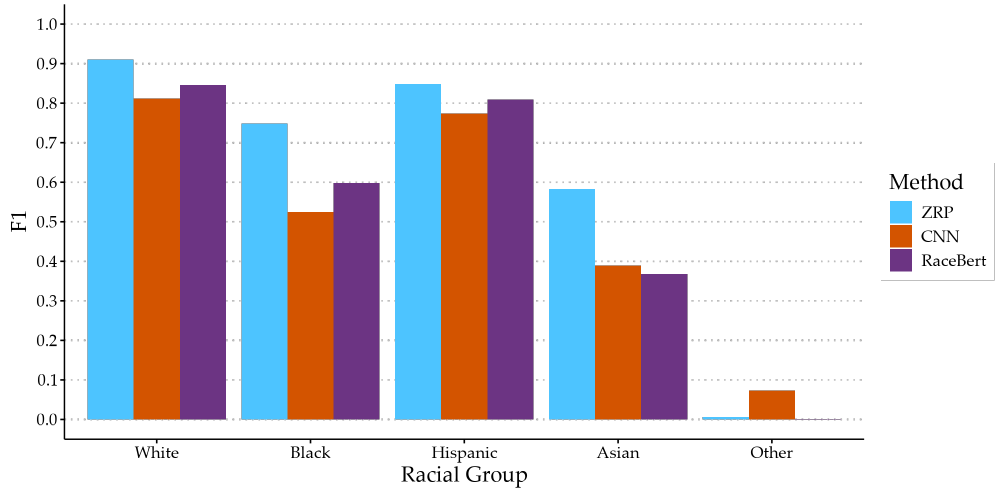
imputation methods. Overall, ZRP consistently outperforms both CNN and RaceBERT for nearly all racial categories, with the exception of "Other." Note that depending on the state, either CNN or RaceBERT dominates the other method. This should come as no surprise because CNN was trained using North Carolina data while RaceBERT was trained using Florida data.

Figure A.1: F1 Scores by Racial Group and Race Imputation Method

(a) North Carolina




(b) Florida



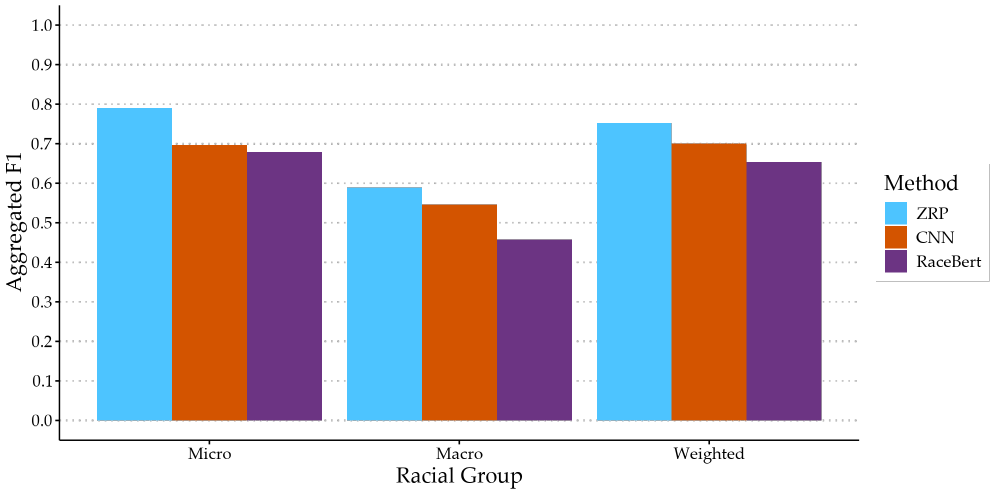

*Notes:* This figure shows the F1 scores across racial groups and race imputation methods using the NC voter file and the FL voter files.

Figure A.2 shows the aggregated F1 scores. The patterns from the F1 scores across race groups remain when the different aggregation methods are used. Regardless of aggregation method (micro, macro, and weighted), ZRP dominates both CNN and RaceBERT in both voter files. In the North Carolina voter file, CNN performs better than RaceBERT. In the

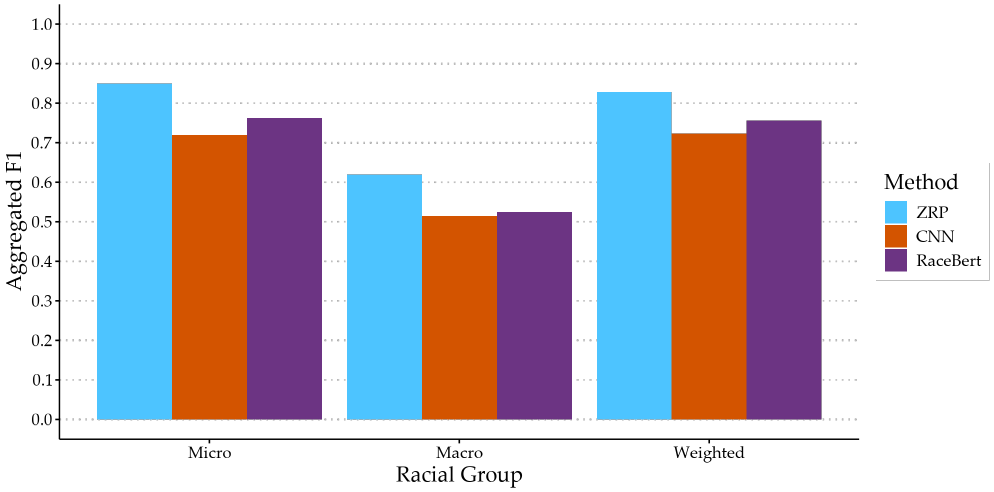
Florida case, the opposite is true.

Figure A.2: Aggregated F1 Scores by Racial Group and Race Imputation Method

(a) North Carolina




(b) Florida



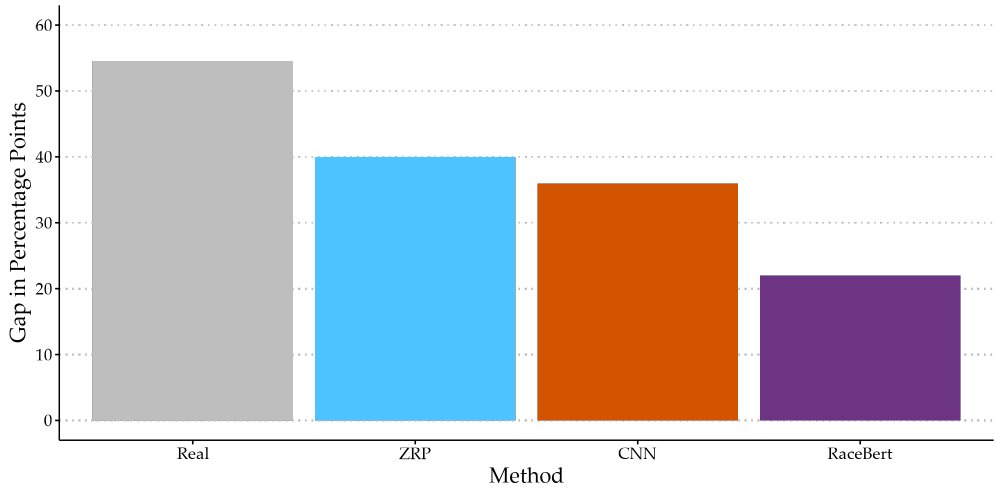

*Notes:* This figure shows the aggregated F1 scores across race imputation methods using the NC voter file and the FL voter files.

Figure A.3 shows the Black-White gap across methods. ZRP is the method that does best at estimating the racial gap, but the difference from the true gap is still large (10-15 percentage points). In the North Carolina voter file, CNN is only a few percentage points away from the gap estimated by ZRP while RaceBERT lags behind. In the Florida voter file, ZRP strikingly outperforms the other two methods – both CNN and RaceBERT estimate gaps that are almost 30 percentage points lower than the true gap observed in the data.

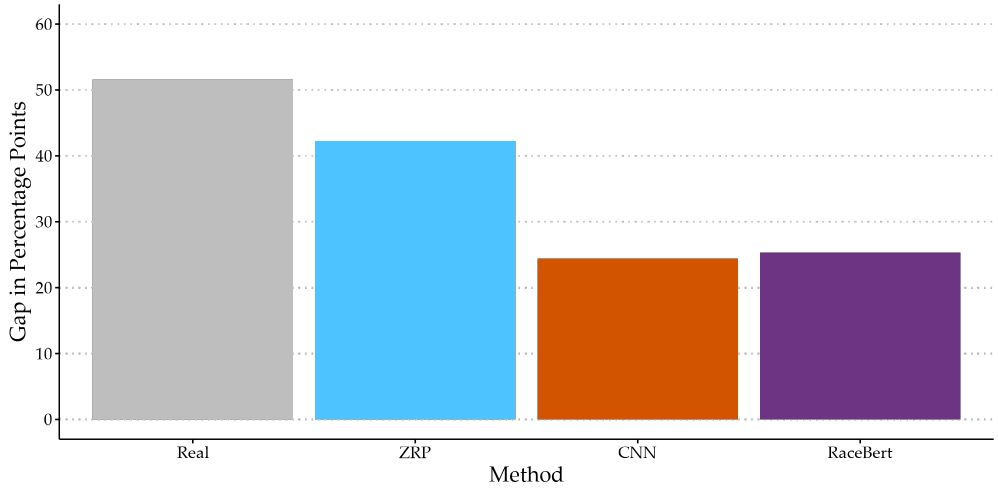
Figure A.4 illustrates the racial gaps for different race groups across the race imputation

Figure A.3: Black-White Gap in Democratic Party Registration by Race Imputation Method

(a) North Carolina




(b) Florida

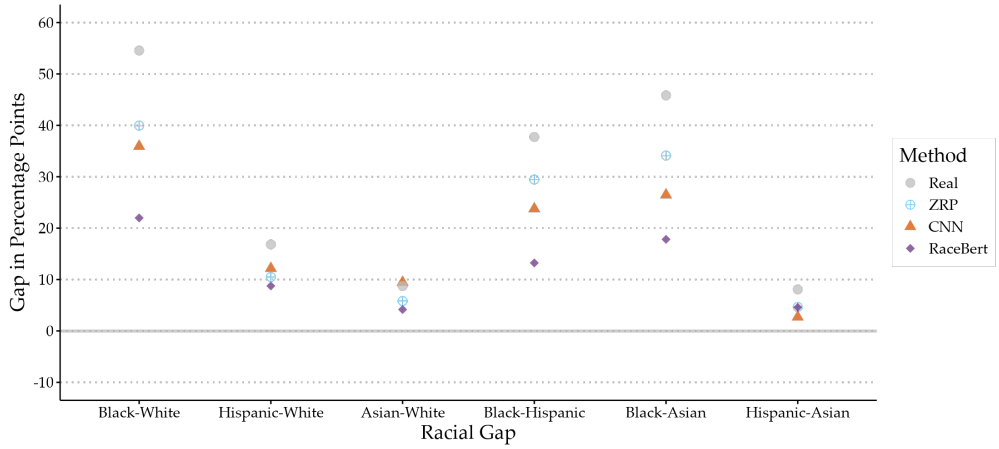



*Notes:* This figure shows the estimated Black-white gap in democratic party registration by different race imputation methods using the NC voter file and the FL voter files.

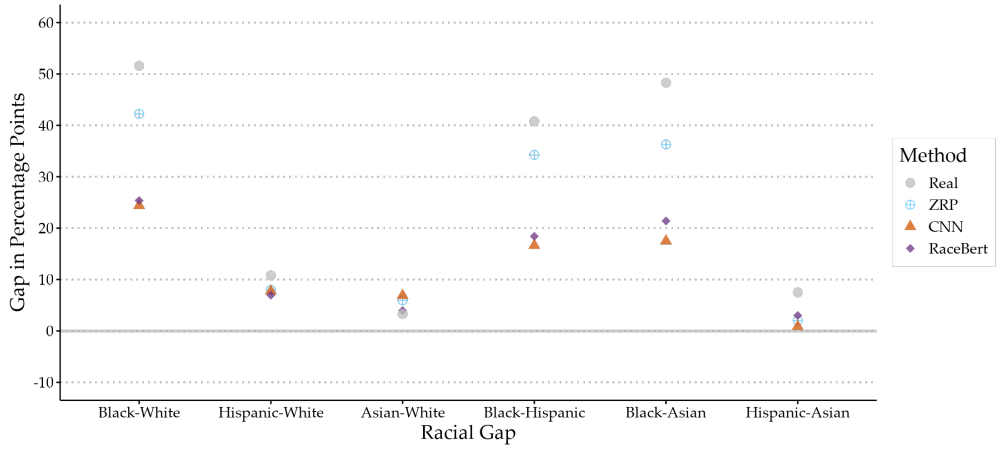
methods. All three methods are far from the real estimated racial gap for most racial groups. For the first three racial gaps (Black-White, Black-Hispanic, Black-Asian), ZRP does the best at estimating the racial gap in both the North Carolina and Florida voter file. However, for Hispanic-White, Hispanic-Asian, and Asian-White gaps, ZRP does not always dominate. (In the North Carolina voter file, ZRP does better for the Hispanic-Asian racial gap, but CNN outperforms for both the Hispanic-White and Asian-White racial gaps. In the Florida voter file, ZRP only dominates in the Hispanic-White racial gap, while RaceBERT does better in the Hispanic-Asian and Asian-White racial gaps.)

Figure A.4: All Racial Gaps in Democratic Party Registration by Race Imputation Method

(a) North Carolina




(b) Florida



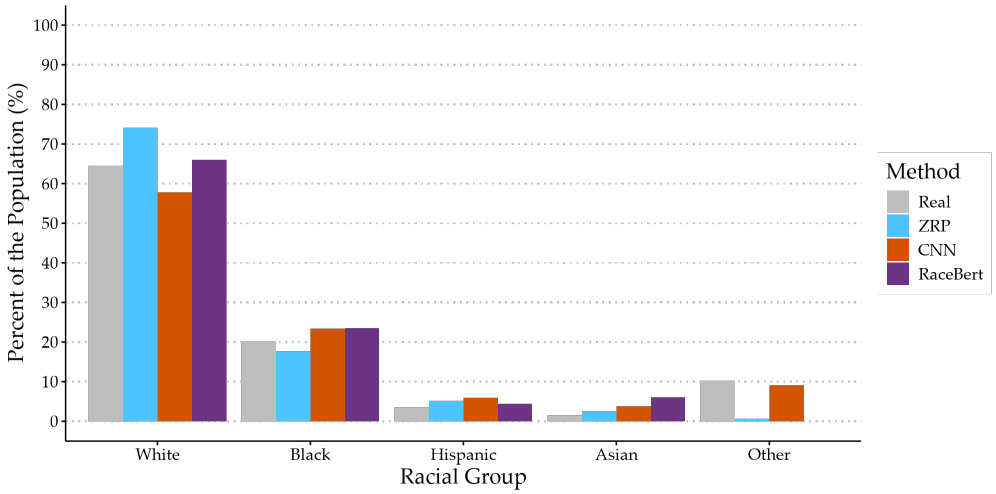

*Notes:* This figure shows the estimated racial gaps in democratic party registration by different race imputation methods using the NC voter file and the FL voter files.

Figure A.5 shows the simulated racial composition across the three imputation methods. In both voter files, ZRP overestimates the percent of white people in the population, while

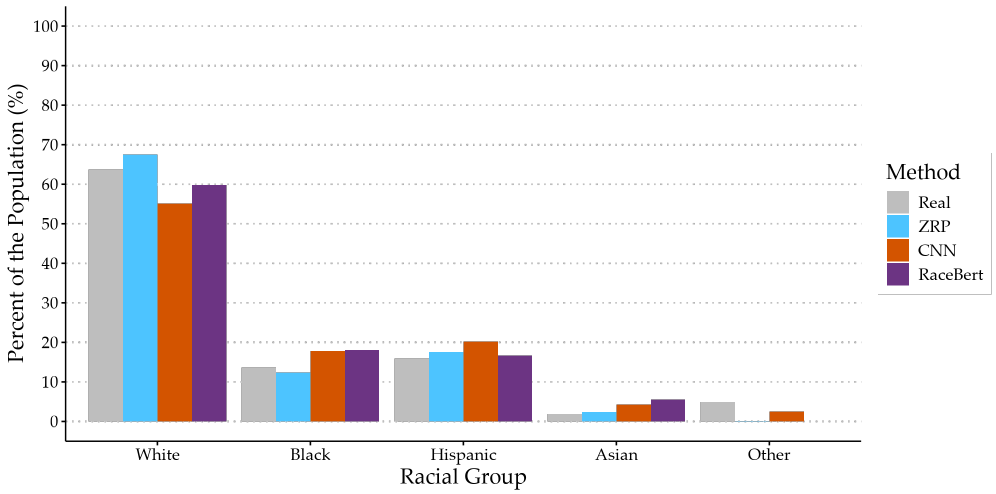
both CNN and RaceBERT overestimate the percent of Black people in the population. Notably, CNN is the only method that reasonably captures the size of the population identifying outside the four major racial categories.

Figure A.5: Racial Composition by Race Imputation Method

(a) North Carolina




(b) Florida




*Notes:* This figure shows the simulated racial composition of the NC and FL voter files under different race imputations.

Tables A.1 and A.2 report distance metrics between the estimated and actual population distributions for the North Carolina and Florida voter files, respectively. For the North Carolina voter file, CNN is the method that comes closer to the real distribution of the population (distance metric of 8.25) followed by RaceBERT and then by ZRP. The pattern is reversed in the Florida voter file. ZRP does best at recreating the population distribution

with a distance metric of 6.43, followed by RaceBERT with a distance metric of 8.59, and finally with CNN with a distance of 11.04. In general, all of the machine learning methods outperform the methods in the main text when matching the distribution of the population for both voter files.

Table A.1: Euclidean Distance between Simulated and Real Demographics across Machine Learning Race Imputation Methods (North Carolina)

| Method | % White | % Black | % Hispanic | % Asian | % Other | Distance Metric |
|---|---|---|---|---|---|---|
| **Real** | 64.52 | 20.15 | 3.57 | 1.51 | 10.24 | NA |
| **CNN** | 57.79 | 23.42 | 5.96 | 3.73 | 9.10 | 8.25 |
| **ZRP** | 74.08 | 17.66 | 5.22 | 2.47 | 0.57 | 13.96 |
| **RaceBert** | 65.98 | 23.52 | 4.47 | 6.01 | 0.02 | 11.79 |

*Notes:* This table shows the simulated racial demographics by race imputation method in the NC voter file data using machine learning approaches. The final column shows the Euclidean distance (in 5-space) between the simulated demographics and the observed demographics in the data.

Table A.2: Euclidean Distance between Simulated and Real Demographics across Machine Learning Race Imputation Methods (Florida)

| Method | % White | % Black | % Hispanic | % Asian | % Other | Distance Metric |
|---|---|---|---|---|---|---|
| **Real** | 63.76 | 13.59 | 15.92 | 1.84 | 4.88 | NA |
| **CNN** | 55.13 | 17.79 | 20.21 | 4.30 | 2.56 | 11.04 |
| **ZRP** | 67.49 | 12.38 | 17.60 | 2.43 | 0.09 | 6.43 |
| **RaceBert** | 59.75 | 18.02 | 16.66 | 5.56 | 0.01 | 8.59 |

*Notes:* This table shows the simulated racial demographics by race imputation method in the FL voter file data using machine learning approaches. The final column shows the Euclidean distance (in 5-space) between the simulated demographics and the observed demographics in the data.

## A.5 Imputation Methods that Require More Inputs

This appendix examines two imputation methods that require supplementary inputs beyond name and geography data. Both methods discussed here address limitations in standard BISG approaches but at the cost of increased data requirements. We exclude these methods from the main text because of their additional data needs.[21] Both methods address similar BISG issues as BIRDiE, which we study in the main text and does not require additional data beyond name and geography.

### A.5.1 Greengard and Gelman (2024)

BISG is a popular imputation method because it uses publicly available Census data and is mathematically straightforward. However, BISG relies on a fundamental assumption that surname and location are conditionally independent of race in the US. Greengard and Gelman (2024) demonstrate that this independence assumption results in systematic inaccuracies, and propose a modification to BISG predictions.

Greengard and Gelman (2024) modify standard BISG predictions through a two-step process. First, they calculate conventional BISG predictions for all individuals in a dataset. Second, they apply raking – an algorithm for adjusting data to match known population margins – to align these predictions with verified demographic distributions. In practice, this requires two new data sources: (1) "the joint distribution of surname and geolocation obtained from state voter registration lists" and (2) "the statewide race/ethnicity distribution of registered voters from the Voter Supplement of the Current Population Survey" (Greengard and Gelman, 2024, p. 2).

For registered voters, necessary margins can be obtained from publicly available data. However, the necessary data for raking may be unavailable in different empirical applications. This method is only applicable when population margins are already known.

### A.5.2 Argyle and Barber (2022)

Argyle and Barber (2022) demonstrate that BISG produces misclassification errors that are strongly correlated with demographic and socioeconomic factors. For instance, Black people who live in high-income areas are more likely to be misclassified as White than Black people who live in low-income areas. To reduce this bias in misclassification errors, the authors propose an ensemble machine learning approach.

[21] Moreover, accessible replication packages are not available for one of the methods.

In particular, they incorporate the standard BISG probabilities for each racial group as inputs into a random forest model with 16 other individual-level or geographic-level political and socioeconomic predictors. The random forest model adjusts the BISG predictions so they are less correlated with political and economic factors.

In practice, they train and tune their random forest model using randomly selected 80% subset of the Florida voter file.[22] Then, they test the random forest model on the remaining 20% of the Florida voter file as well as the North Carolina voter file.

On top of the standard BISG predictions, Argyle and Barber (2022)'s random forest uses 16 additional variables from voter files. Their pre-trained random forest model is only usable when researchers have access to the same inputs. That said, researchers hypothetically could use their code to train their own random forest models to include only geography-level variables (which are knowable conditional on having geography data). This exercise is out of scope for this paper, but could provide more generalizable random forest models for practitioners.[23]

[22] Their replication documentation explains that the training requires significant computational capacity: "this code is set up to run in parallel on 16 processors, and that the original run took 48+ hours." (Barber and Argyle, 2023).

[23] This is out of scope because our aim is to compare pre-existing methods.

## A.6 Black Box Private Sector Methods

### A.6.1 Imputation in Purchased Data

Empirical researchers are increasingly using private data for research (Einav and Levin, 2014). Proprietary datasets that include race variables may be derived from their own proprietary race imputation methods.

A prominent example is L2's national voter file, which has been used in researching bankruptcy (Argyle et al., 2025), the Vietnam war (Bleemer, 2025), political identity (Frake et al., 2024), and more. L2 imputes race unless it is self-reported in the underlying state voter file (Frake et al., 2024; Bleemer, 2025). Their imputation is proprietary (Argyle et al., 2025), but researchers have linked L2 to other data (such as SNAP records or college applications) and documented discrepancies between L2's imputed race and self-reported race in those administrative sources (Chyn and Haggag, 2023; Bleemer, 2025).

### A.6.2 NameSor

NamSor is a private company that sells race/ethnicity imputation (NamSor, 2025c). NamSor advertises that they can provide the origin, ethnicity, the US race, country, and gender of names (NamSor, 2025b).

The tool has been used in 364 research papers (as of February 2025) according to their website (NamSor, 2025c). Most papers that use NamSor are in the medical community or are papers that analyze academic publications biases. To the best of our knowledge, Sebo (2023) is the only paper to evaluate the performance of NamSor against data with known identification. That said, there has not been a comparative evaluation of NamSor against other methods like BISG or NamePrism.[24]

In this draft of the paper, we do not compare NamSor against other methods. We make this choice for two reasons. First, NamSor is prohibitively expensive for many research contexts. NamSor's pricing page (as of February 2025) shows that it costs $4,167 to process 500,000 names (6% of the full NC voter file sample) and $14,000 to process 2,000,000 names (25% of the full NC voter file) (NamSor, 2025a). Second, the methods used in NamSor are not transparent. For illustration, Sebo (2023) states that "The methodology used by the algorithm to determine the most likely origin of individuals is relatively opaque to

[24]NamSor has more in common with NamePrism than other methods in our main text. NamSor and NamePrism are unique in that they have international use cases.

non-specialists, but likely relies on large databases combining names with cultural, ethnic and linguistic backgrounds."

In this paper, we compare free methods for race imputation with technical documentation. We do not extend the exercise to include black box tools for purchase.

## A.7 Results using Florida Voter File

In the main text, we use a 500,000 observation random sample of the 2023 North Carolina voter file to test performance. One fundamental concern with this approach is that the patterns and results might differ if we used data from a different place. To address this concern, we present our main results in comparison to results generated from using a 500,000 observation random sample of a different voter file – the Florida 2017 voter file (Sood, 2022).[25] Overall, we show below that the key patterns from the analysis using the North Carolina still hold in the Florida data.

Figure A.6 compares F1 scores by racial group across the main race imputation methods for the Florida voter files. The main text results showed ZRP consistently outperforms the other imputation methods for nearly all racial categories, with the exception of "Other" in the North Carolina voter file. The same pattern emerges in the Florida voter file where ZRP dominates the other methods. Note that the zip method does better for the Hispanic population in the Florida voter file compared to the North Carolina voter file. This is because there are very few zips in North Carolina where the most common race is Hispanic, so this measure performs poorly. In Florida, since there are more Hispanic people and Hispanic-heavy zip codes, this is less of an issue.

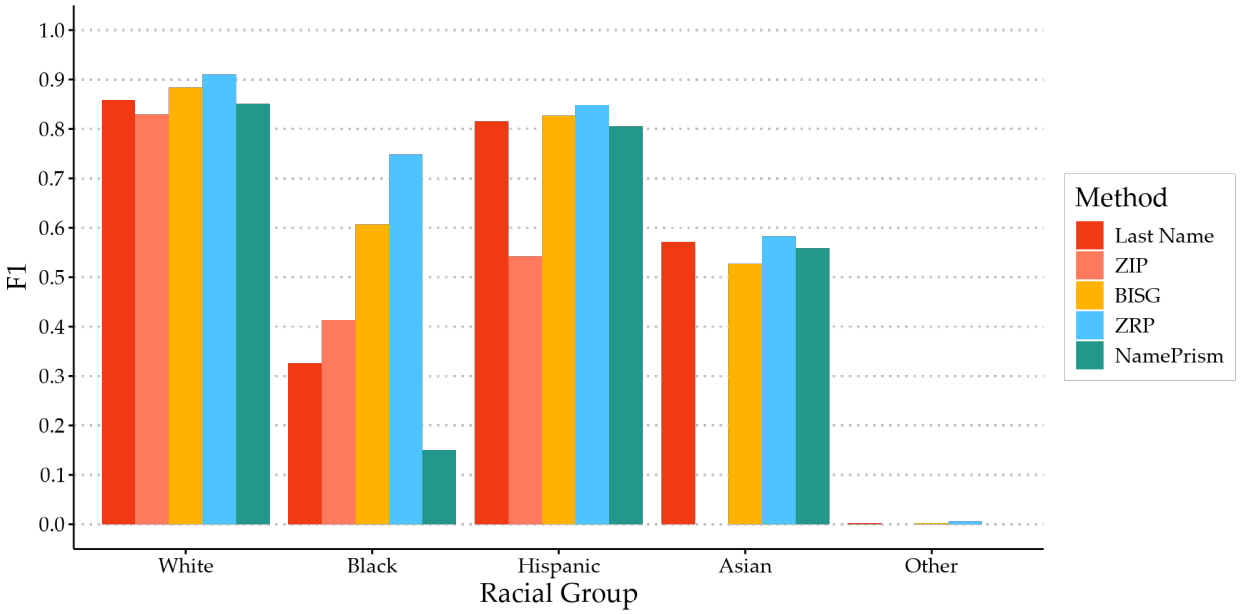
Figure A.7 shows the aggregated F1 scores. The patterns from the F1 scores across race groups remain when the different aggregation methods are used. For all aggregation methods (micro, macro, and weighted), ZRP dominates other methods in both voter files. Overall, methods perform similarly relative to one another in both states.

Figure A.8 shows the Black-White gap in democratic registration in the Florida voter file across methods. As was in the case in the North Carolina voter file, the BIRDiE method does the best in Florida too. After BIRDiE, ZRP is the method that does best in both the North Carolina and Florida voter files. In both cases, however, the difference between the ZRP gap and the observed gap remains large. The gap between the observed gap and the gap estimated by ZRP is about 15 percentage points in the North Carolina context and about 10 percentage points in the Florida context. All methods perform In the North Carolina voter file, all methods perform worse relative to the identified gap in the Florida voter file.

Figure A.9 illustrates the racial gaps for different pairs of race groups across the race imputation methods in the Florida data. As was the case in North Carolina, BIRDiE

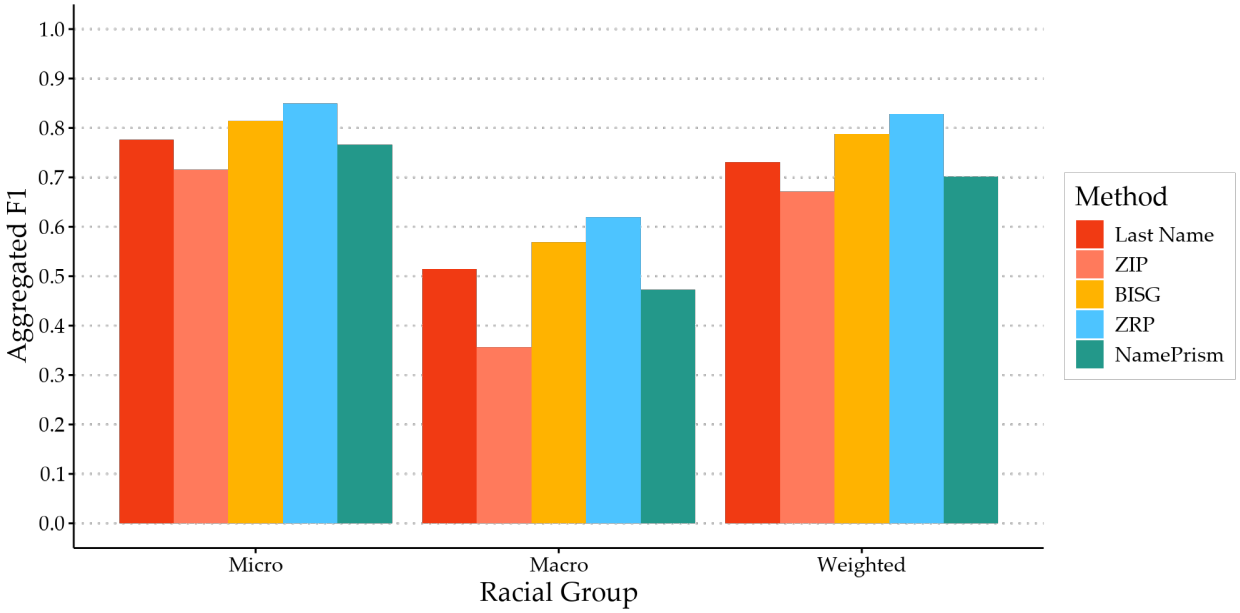
[25] This Florida voter file dataset has been used in previous papers such as Chintalapati et al. (2018) and Parasurama (2021).

Figure A.6: F1 Scores by Racial Group and Race Imputation Method (Florida)



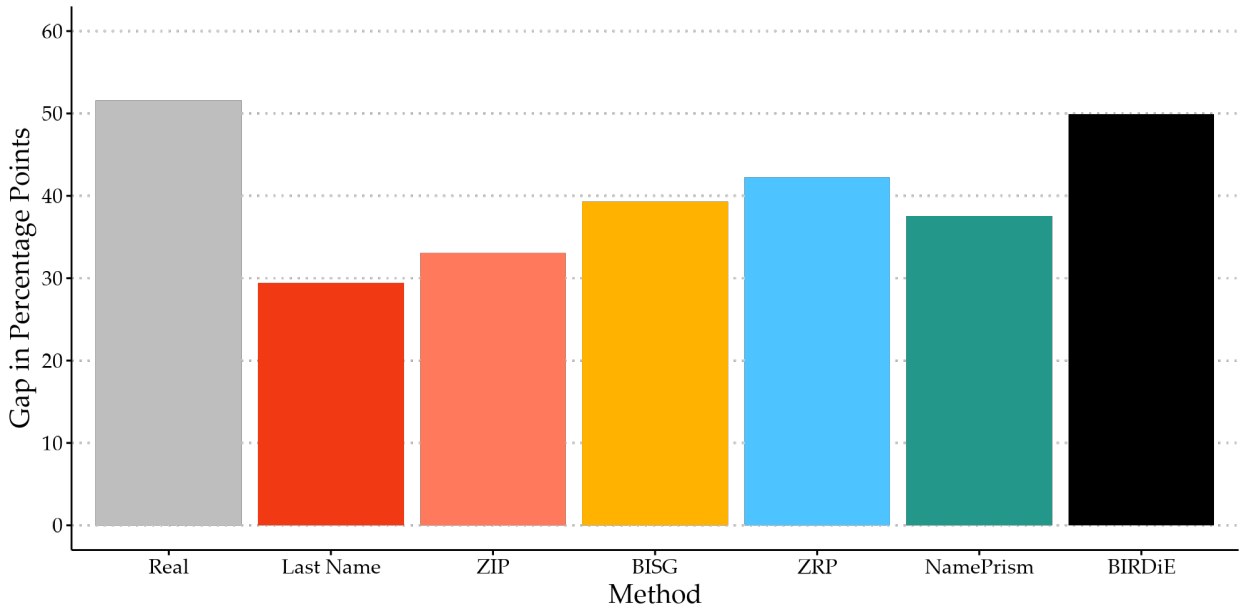

*Notes:* This figure shows the F1 scores across racial groups and race imputation methods using the Florida voter file.

Figure A.7: Aggregated F1 Scores by Racial Group and Race Imputation Method




*Notes:* This figure shows the aggregated F1 scores across race imputation methods using the Florida voter file.

Figure A.8: Black-White Gap in Democratic Party Registration by Race Imputation Method



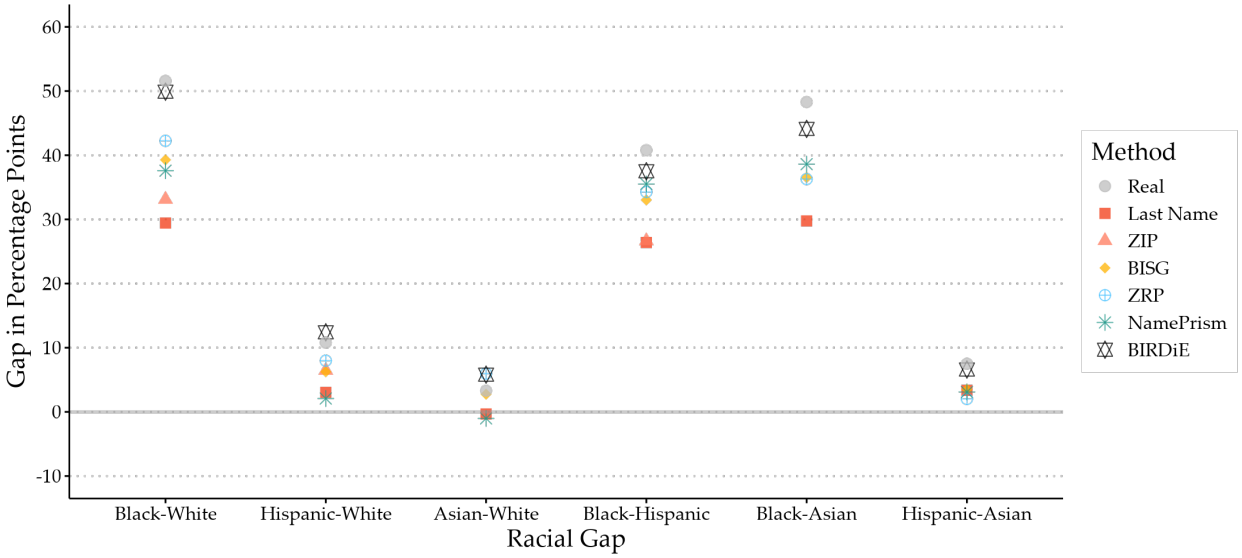

*Notes:* This figure shows the estimated Black-white gap in democratic party registration by different race imputation methods using the Florida voter file.

outperforms in most racial gaps over the other methods in both the North Carolina and the Florida voter file. The three main take-aways from the main text analysis using North Carolina data hold in the Florida case: disparity estimates can go in the wrong direction, imputation methods are not rank-preserving across different racial gaps of interest, and disparity estimates can be far from the true values.

Figure A.10 shows the racial composition for all the imputation methods in Florida. The main demographic difference between NC and FL is that there is a higher percentage of Hispanic people in Florida and a higher percentage of Black people in NC. In both voter files, all methods overestimate the share of the white population and underestimate the share of the Black population.

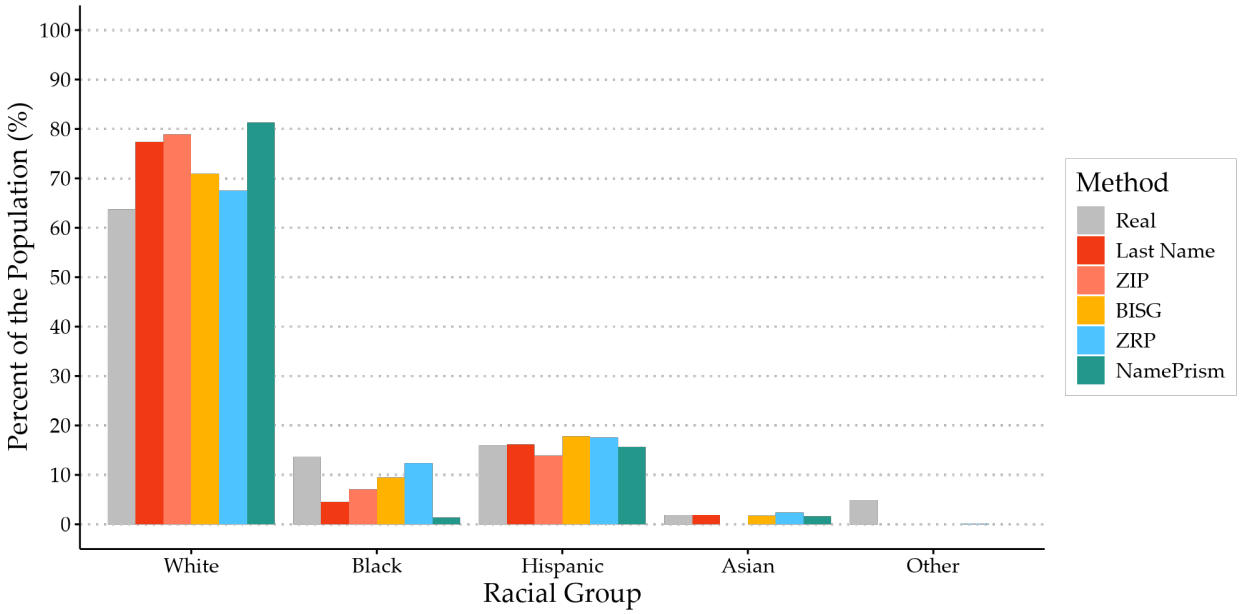
Table A.3 reports distance metrics between the estimated and actual population distributions in the Florida data. For both North Carolina and Florida, ZRP is the method that comes the closest to the real distribution of the population (with a distance of 13.96 in the North Carolina voter file and 6.43 in the Florida voter file). BISG follows with a distance of 18.53 and 9.82 for the North Carolina and Florida voter files respectively. NamePrism is the furthest from the observed distribution in both cases. There is one distinction between North Carolina and Florida in the methods rankings: zip code performs better than last

Figure A.9: All Racial Gaps in Democratic Party Registration by Race Imputation Method




*Notes:* This figure shows the estimated racial gaps in democratic party registration by different race imputation methods using the Florida voter file.

Figure A.10: Racial Composition by Race Imputation Method




*Notes:* This figure shows the simulated racial composition of the Florida voter file under different race imputations.

name in North Carolina, but the reverse is true in Florida. This is explained by the fact that Florida has a larger share Hispanic population (for whom last name outperforms zip), and North Carolina has a larger share Black population (for whom zip outperforms last name).

Table A.3: Euclidean Distance between Simulated and Real Demographics across Race Imputation Methods (Florida)

| Method | % White | % Black | % Hispanic | % Asian | % Other | Distance Metric |
|---|---|---|---|---|---|---|
| **Real** | 63.76 | 13.59 | 15.92 | 1.84 | 4.88 | NA |
| **Last Name** | 77.33 | 4.57 | 16.13 | 1.94 | 0.02 | 17.00 |
| **ZIP** | 78.92 | 7.13 | 13.95 | 0.00 | 0.00 | 17.39 |
| **BISG** | 70.96 | 9.44 | 17.84 | 1.73 | 0.03 | 9.82 |
| **ZRP** | 67.49 | 12.38 | 17.60 | 2.43 | 0.09 | 6.43 |
| **NamePrism** | 81.34 | 1.32 | 15.69 | 1.65 | 0.00 | 21.99 |

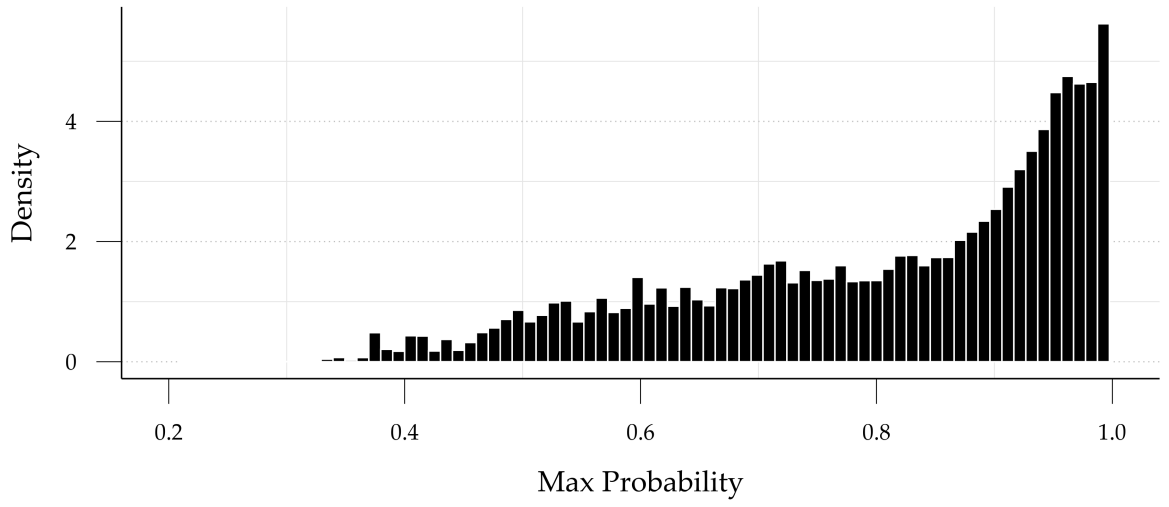
*Notes:* This table shows the simulated racial demographics by race imputation method in the FL voter file data. The final column shows the Euclidean distance (in 5-space) between the simulated demographics and the observed demographics in the data.

## A.8 Comparing Results across Thresholding Options

All race imputation methods generate predicted probabilities that individuals belong to one of the five main racial groups (White, Black, Hispanic, Asian, and Other). In the main text, for each imputation method, we assign people to the racial group which received the highest predicted probability. This is the standard procedure in many papers, including Berry et al. (2024), Argyle and Barber (2022), Letzler et al. (2017), and Cohle and Ortega (2023).

Therefore, observations vary in the predicted probabilities that determine imputation. Figure A.11 shows the distribution of the maximum predicted probabilities across all observations in our 500,000 sample – these are the probabilities that determine assignment to a group. The median probability is high at 85.6%, but values can be as low as 22% and still generate assignments.

Figure A.11: Distribution of Maximum Probabilities




*Notes:* In the main text, imputed race is assigned to the racial group with the highest predicted probability. This figure shows the distribution of those predicted probabilities that result in assignment.

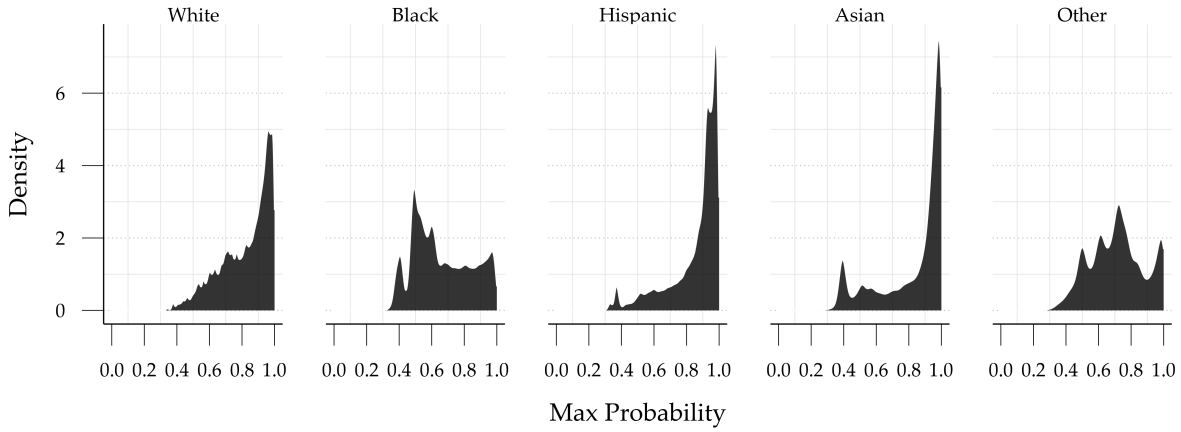
There are alternative ways of assigning individuals to racial groups using the predicted probability output. One common alternative is to assign people to racial groups based on set thresholds.[26] How would such thresholds change the overall imputation results? This

[26]This has been done in previous papers, such as Diamond and Diamond (2024), Luh (2023), Goncalves and Mello (2021), and Feigenberg and Miller (2022). Another alternative is to use the generated probability to randomly assign someone to a racial group. Greenwald et al. (2024) use this procedure for assigning

appendix section explores this question by comparing our baseline results to alternative results using thresholds of increasing strictness – 50%, 70%, and 90%.

Before comparing results across threshold options, we first show motivational evidence that thresholding will not equally impact distinct racial groups. Figure A.12 demonstrates the distribution of predicted probabilities split out by the baseline assigned race.

Figure A.12: Density of Maximum Probabilities by Imputed Race




*Notes:* In the main text, imputed race is assigned to the racial group with the highest predicted probability. This figure shows the density plots of the predicted probabilities that result in assignment, faceted by imputed racial group.

The distributions for White, Hispanic, and Asian individuals are right-skewed – the median probabilities for these groups are 87%, 92%, 94%, respectively. In contrast, the distributions for Black and Other racial groups are noticeably flatter and more dispersed. The median probabilities for Black and Other are 64% and 72%, respectively. The median probabilities for these two groups are below the 25th percentile probabilities for the other three groups.

Figure A.12 necessarily implies that thresholding at higher probability levels will disproportionately reduce the number of individuals predicted to be Black (and Other) relative to other groups. However, it is ambiguous how thresholding would impact F1 scores or racial gaps across groups.

We now consider how thresholding impacts our results across multiple thresholds: 50%, 70%, 90%. We compare these three thresholds with the baseline method of assigning to the most common group. Mimicking the structure in the main text, we demonstrate the

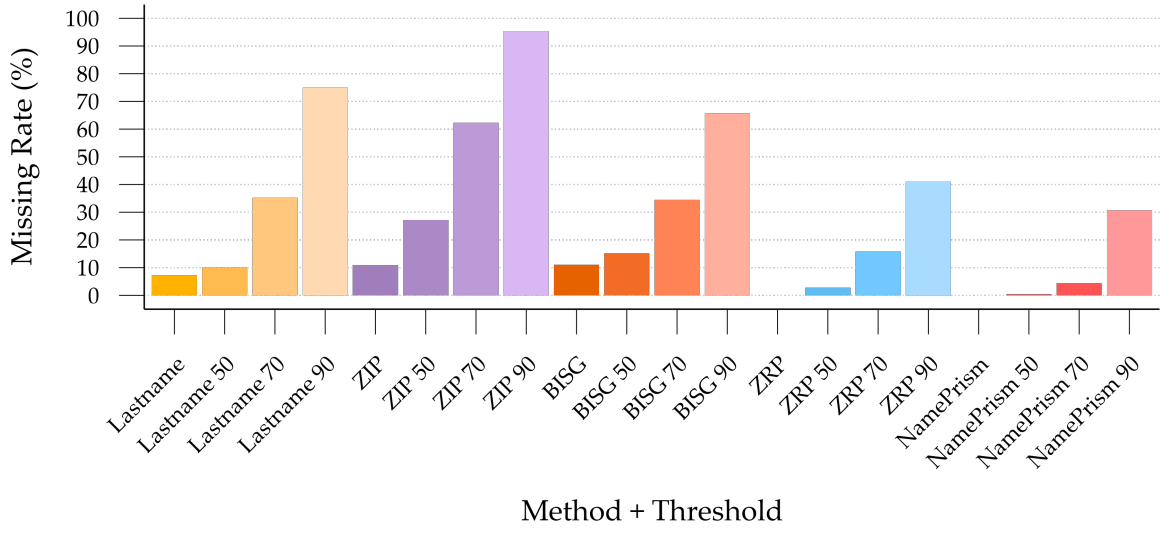
people as Black or non-Black using BISG-generated probabilities.

effect of these thresholds on: missing rates, F1 scores, racial gap estimates, and population descriptives.

### A.8.1 How does thresholding impact missing rates?

Fig A.13 shows that, as the threshold becomes more stringent, missing rates mechanically increase for all imputation methods. The Zip method is the one most affected – if the threshold is set at 90%, more than 80% of the observations are missing race.[27] Meanwhile, NamePrism is the method least affected by strict thresholds. If the threshold is set at 90%, only 30% of observations are missing race.

Figure A.13: Missing Rate by Methods and Thresholds




*Notes:* This figure shows the percent of observations without a race prediction by the main text methods and the threshold options. Threshold options go from least restrictive to most restrictive as follows: baseline (main text option), 50% threshold, 70% threshold, and 90% threshold. Bars are colored by imputation method and shaded based on the threshold option.

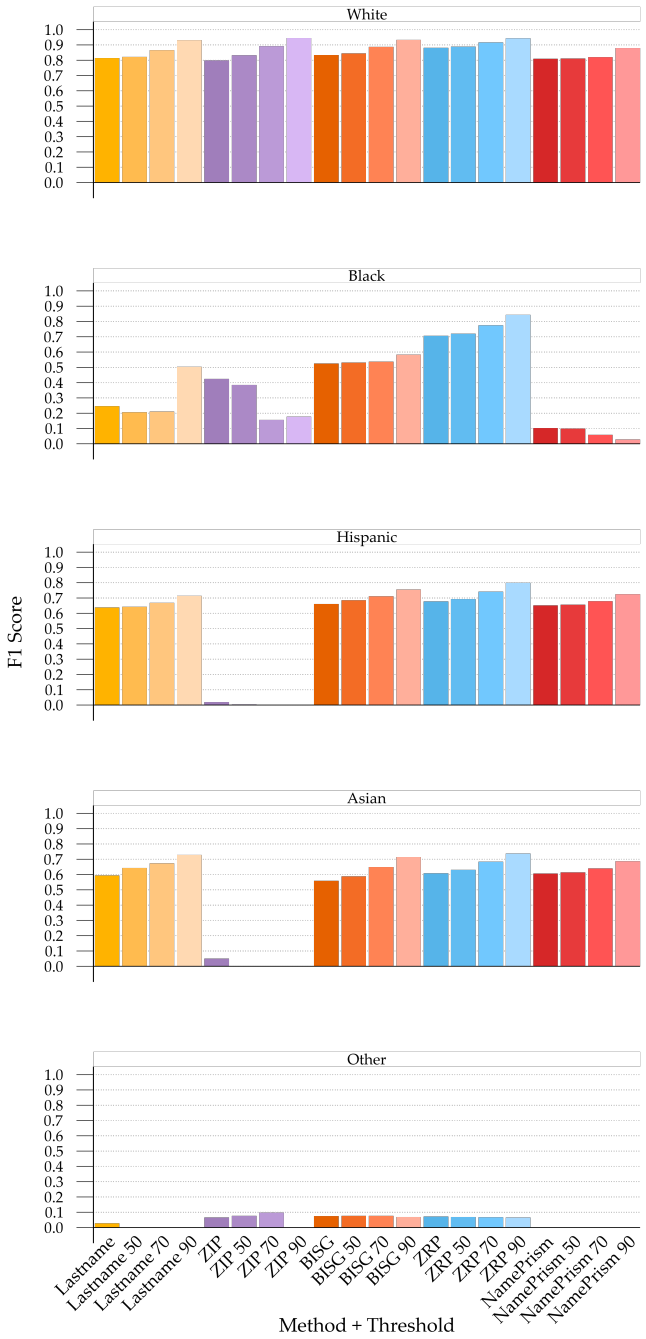
### A.8.2 How does thresholding impact F1 scores?

Figure A.14 demonstrates that when thresholds are increased, F1 scores improve for most racial groups and methods. However, there are some exceptions for the Black population. When using zip imputation and NamePrism imputation, setting a higher threshold actually

[27]The baseline Zip method has a 10 percent missing rate due to data limitations in the ACS. For ZIP codes with small populations, race distributions are not published in order to protect individual privacy.

decreases F1 scores. Therefore, it is not universally true that setting high thresholds will improve recall and precision for all racial groups.

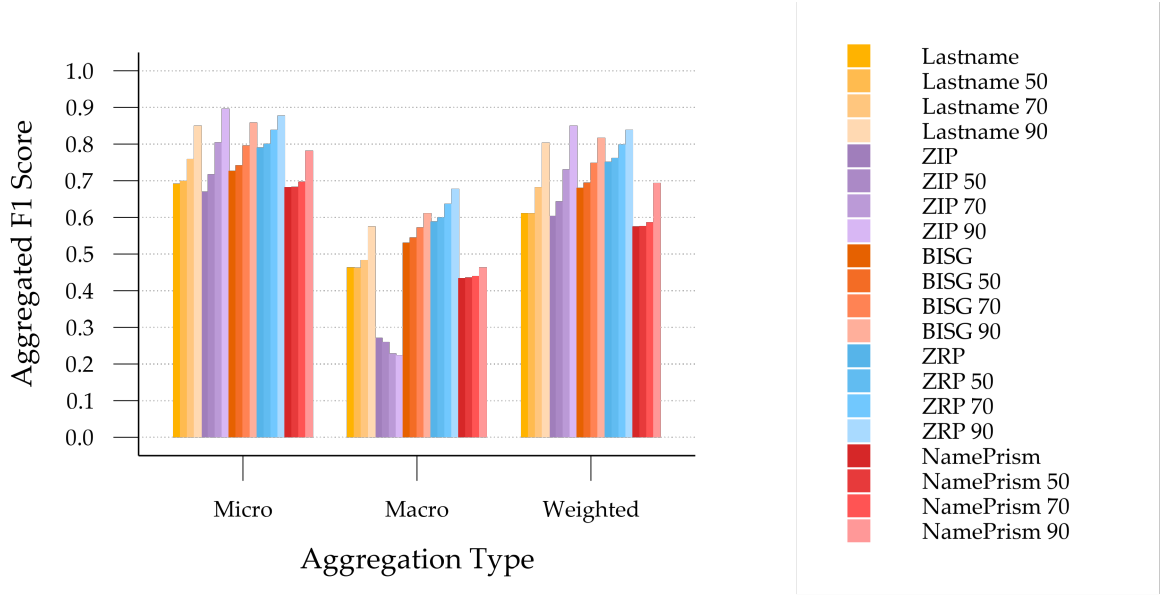
Figure A.14: F1 Scores by Race, Methods, and Thresholds




*Notes:* This figure shows the F1 scores by the main text methods and the threshold options, faceted by racial group. Threshold options go from least restrictive to most restrictive as follows: baseline (main text option), 50% threshold, 70% threshold, and 90% threshold. Bars are colored by imputation method and shaded based on the threshold option.

Figure A.15 demonstrates how aggregated F1 scores compare across thresholds options and imputation methods. Consistent with higher thresholds improving F1 scores for most racial groups and most methods, most aggregated F1 scores increase as thresholds increase. The one exception is the macro aggregation for zip – this happens because that averaging method puts enough weight on the lower F1s for the Black population that it offsets the high F1s for the White population.

Figure A.15: Aggregated F1 Scores by Methods and Thresholds

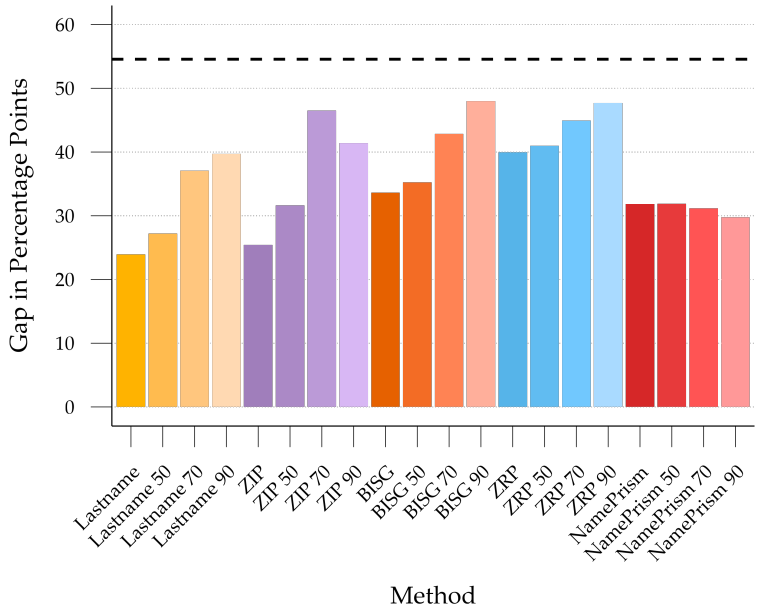



*Notes:* This figure shows the aggregated F1 scores by the main text methods and the threshold options. There are three types of aggregation (as in the main text): micro, macro, and weighted. Threshold options go from least restrictive to most restrictive as follows: baseline (main text option), 50% threshold, 70% threshold, and 90% threshold. Bars are colored by imputation method and shaded based on the threshold option.

### A.8.3 How does thresholding impact racial gap estimation?

Fig A.16 shows how the estimated Black-White racial gap in democratic registration varies as you change the threshold across methods. As the threshold increases, the gap estimation improves compared to the baseline for ZIP, BISG, ZRP, but this is not true for NamePrism. Even though the zip method with the 90% threshold yielded the most missing race observations and one of the lowest F1 scores for the Black population, this method comes the closest to replicating the observed Black-White gap in the data.

Figure A.16: Black-White Gap in Democratic Party Registration by Method and Threshold

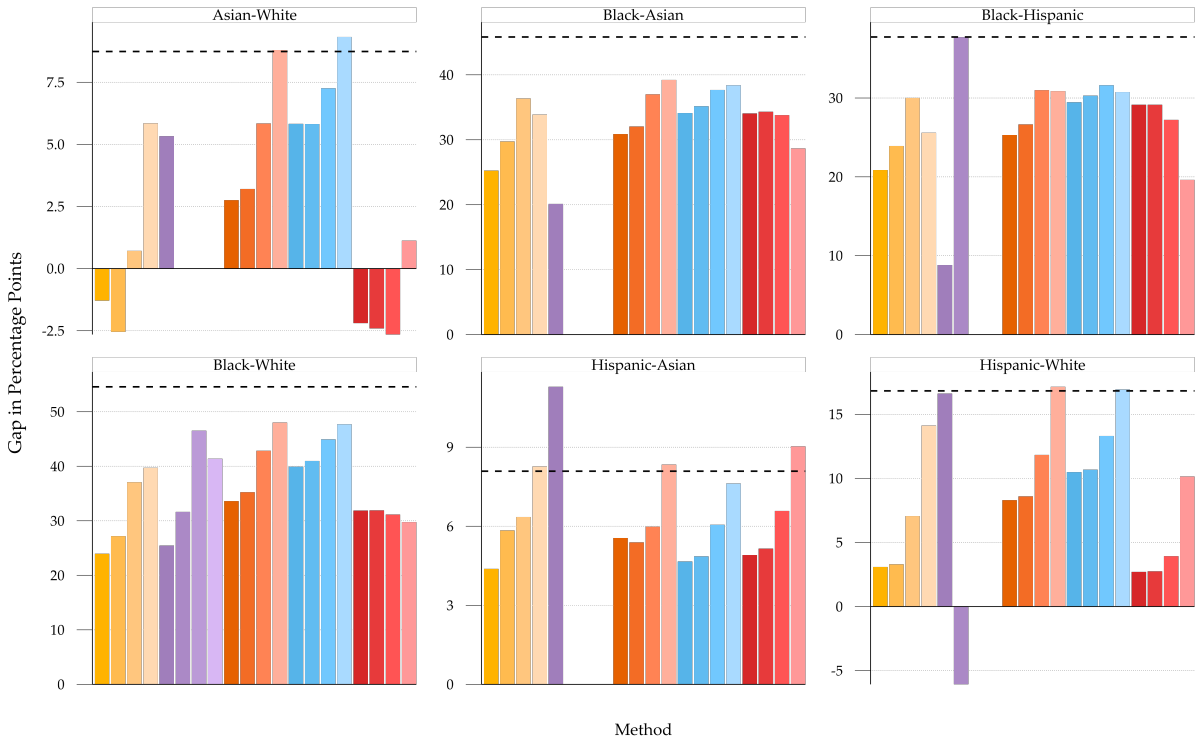



*Notes:* This figure shows the estimated Black-White disparity in democratic party registration by different race imputation methods. The dashed black line is the observed gap in the data. Threshold options go from least restrictive to most restrictive as follows: baseline (main text option), 50% threshold, 70% threshold, and 90% threshold. Bars are colored by imputation method and shaded based on the threshold option.

What about other racial gaps? Now, we use Figure A.17 to compare all possible racial gaps across the main groups (White, Black, Hispanic, Asian) across the 5 methods (last name, ZIP, BISG, ZRP, NamePrism) and 4 threshold options (baseline, 50%, 70%, 90%).

Figure A.17 shows the distribution of racial gaps across six pairwise race comparisons: Black-White, Hispanic-White, Asian-White, Black-Hispanic, Black-Asian, and Hispanic-Asian. In the case of two gaps, the Hispanic-White and Hispanic-Asian gaps, higher thresholds yield more accurate racial gap estimates regardless of method. But, this is not uniformly true for the other four gaps: Black-White, Asian-White, Black-Hispanic, and Black-Asian. Across the 25 gap and method combinations with results (excluding various combinations with zip), higher thresholds improve estimation 16 times. Higher thresholds improve estimation more consistently for ZRP or BISG than for surname imputation or NamePrism.

Figure A.17: Racial Gaps in Democratic Party Registration by Method and Threshold

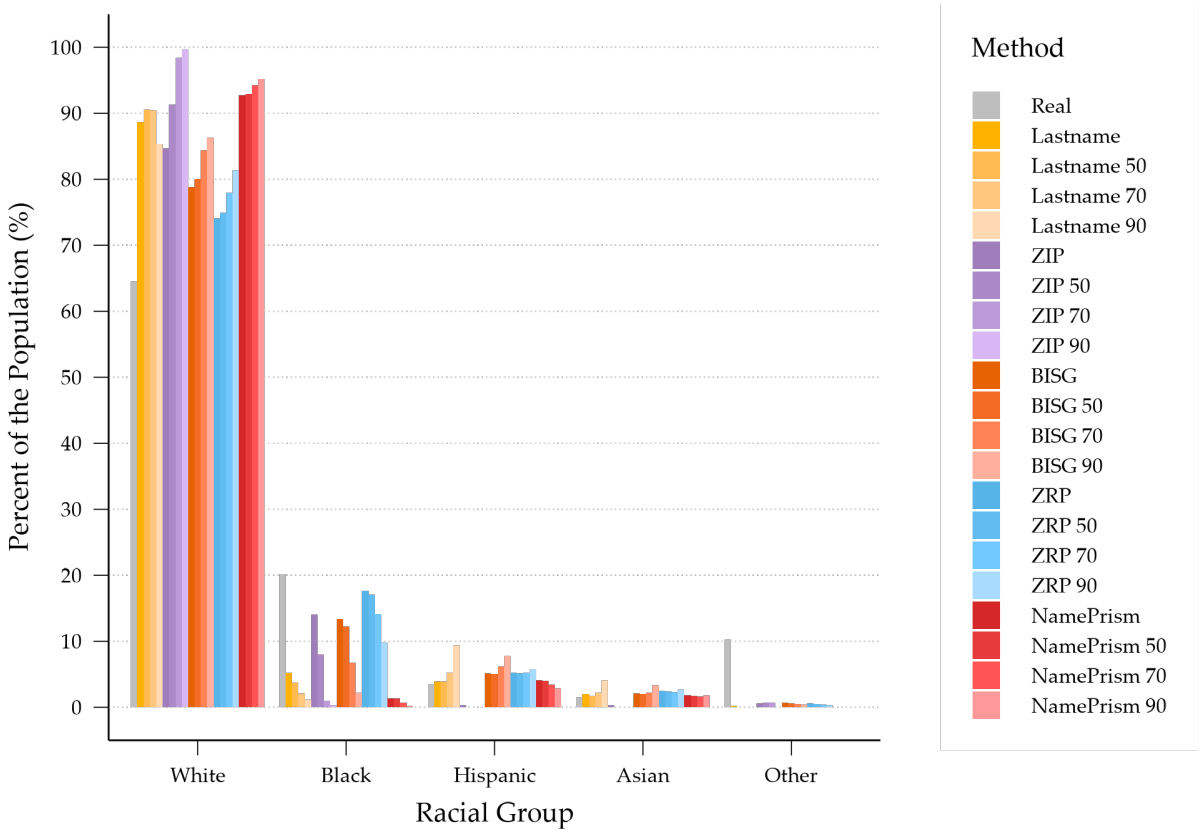



*Notes:* This figure shows estimated racial disparities in democratic party registration by different race imputation methods across six gaps of interest: Black-White, Hispanic-White, Asian-White, Black-Hispanic, Black-Asian, and Hispanic-Asian. The dashed black line is the observed gap in the data. Threshold options go from least restrictive to most restrictive as follows: baseline (main text option), 50% threshold, 70% threshold, and 90% threshold. Bars are colored by imputation method and shaded based on the threshold option.

### A.8.4 How does thresholding impact population descriptives?

Figure A.18 shows how imputed population descriptives vary across threshold options. For most methods, when the threshold increases, the imputed population becomes a higher percent White, moving all methods further away from the actual observed rates of White people in the population.[28] Symmetrically, the non-white imputed population generally declines as thresholds increase and diverges further from the non-white true rate in the population.

Figure A.18: Racial Composition by Method and Threshold




*Notes:* This figure compares the racial compositions of the sample under different race imputation methods and threshold options. The black bar is the observed rate for a group in the data. Threshold options go from least restrictive to most restrictive as follows: baseline (main text option), 50% threshold, 70% threshold, and 90% threshold. Bars are colored by imputation method and shaded based on the threshold option.

Figure A.18 shows that as thresholds increase, the Black share of the imputed population decreases, veering further from the actual rate. In contrast, the Hispanic share declines

[28] One exception is the 90% threshold for last name, which reduces the percent white – this is likely because of the distributions shown in Fig A.12. Moving from the critical threshold of 70% to 90% loses more white than non-white predictions because of the right-skewed prediction distributions for Hispanic and Asian people.

slightly for ZIP, ZRP, and NamePrism, indicating that stricter thresholds may filter out individuals otherwise classified as Hispanic by these approaches. For the Asian share, the ZIP method assign zero population to thresholds different to using the baseline, highlighting a clear limitation of ZIP code data for identifying Asian individuals. In contrast, for all other methods, the Asian share increases as the threshold becomes more stringent. As expected, all thresholding methods put the rate of people imputed as Other very close to 0.

### A.8.5 Take-aways

There are a few unambiguous effects of setting higher thresholds when imputing race. For one, increasing thresholds mechanically increases the missing rate because it mandates dropping more observations. Because the prediction distribution is flatter for Black people than other groups (Fig A.12), higher thresholds means reducing the population imputed as Black (Fig A.18). Generally higher thresholds also lead to higher overestimates of the White population (Fig A.18). (One exception arises at the high-end of thresholds for surname imputation – moving from a 70% to 90% threshold decreases the White population because of the higher predictions for the Asian and Hispanic populations.)

The effects of higher thresholds on racial gaps and individual-level performance are mixed depending on the imputation method and racial groups. Setting higher thresholds increases F1 scores for most methods and racial groups, but this is not always true. F1 scores actually decrease for the Black group as thresholds become stricter when using NamePrism or zip imputation. Likewise, higher thresholds improve estimating racial gaps in some cases but not others. High thresholds tend to improve estimation of all racial gaps for BISG and ZRP, but higher thresholds often make estimation worse when imputing with NamePrism or surnames.

Importantly, these results again highlight the disconnect between individual-level prediction (F1 scores) and racial gaps. A salient example: setting higher thresholds for zip imputation reduces F1 scores for the Black population, but improves Black-White gap estimation. Therefore, as in the main text, improving performance in one dimension through methodological changes does not necessarily improve performance in other dimensions.