Imputing Race

Alex Albright*
Juliana Gamboa-Arbelaez†

February 2025
(Preliminary Draft – Do Not Distribute/Cite)

Abstract

Missing data on race is a fundamental challenge to the empirical study of race in the economy. When race is unobserved, researchers across academia, government agencies, and private companies impute race with other observable variables. There are many ways to impute race, but these methods are under-discussed even though they can have first-order effects on empirical conclusions. In this paper, we outline the current landscape of race imputation methods and compare performance across methods using unique data that includes self-reported race. When evaluating performance, we describe three distinct objects of interest to researchers and practitioners: individual-level classification success, success at estimating racial disparities, and success at estimating racial demographics. We find that the best-performing imputation method depends on the object of interest. No method dominates across all types of performance. This has important research and policy implications since it means practitioners and researchers must consider their end goal when they impute race.

*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, and Jacob Goldin for useful feedback on this project. We thank Kasey Matthews for helping us implement Zest Race Prediction (ZRP). 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

Estimating racial demographics and disparities is important for researchers, policy-makers, and organizations. However, race is often unobserved in datasets, which creates an obvious challenge for empirical study. How can researchers study racial heterogeneity without data on race? How can government agencies conduct equity assessments without data on race? In practice, practitioners *impute* race, meaning they predict racial identity using other observables, such as name and place.

Race imputation happens frequently in academic research, policy-making, and organizational decision-making. In academic research, race imputation has been used to study the racial dimensions of eviction (Hepburn et al., 2020), police brutality (Edwards et al., 2019), labor market misconduct (Egan et al., 2022), pandemic lending programs (Howell et al., 2024), the opioid crisis (Cohle and Ortega, 2023), credit access (Frame et al., 2021), and retirement saving behavior (Coyne et al., 2024). Many US government agencies use race imputation as well: the Consumer Financial Protection Bureau (CFPB) imputes race to evaluate fair lending (Consumer Financial Protection Bureau, 2014), the U.S. Equal Employment Opportunity Commission (EEOC) imputes race to analyze demographic disparities in employment in labor discrimination cases (Harris, 2020), the Department of Health & Human Services (HHS) and the Center for Medicare and Medicaid Services (CMS) impute race to study healthcare disparities (Sorbero et al., 2022; Fremont et al., 2016), and Treasury and the Congressional Budget Office (CBO) impute race to study effects of tax policies on different groups (Fisher, 2023).[1] In the private sector, race imputation has been used by tech companies (e.g., Meta) to assess racial bias in advertisement delivery (Greengard and Gelman, 2024), and national or regional health plans (e.g., Anthem, Cigna, Highmark, and Kaiser) to add race information in their data (Fremont et al., 2016).

Imputing race has become standard practice for dealing with missing data on race across many contexts. While race imputation underlies an enormous amount of empirical analysis, there is limited research on the landscape of imputation methods. ***How does imputation work? What is the current menu of race imputation options? How well do these methods perform (and on what margins of performance)?***

In this paper, we answer these questions in two parts. First, we outline the landscape of race imputation methods, including new methods introduced in the last few years. Second, we compare performance across imputation methods using unique datasets where

[1] Race imputation has also been upheld in the US Court of Appeals for use in federal voting rights cases (Greengard and Gelman, 2024).

individual-level race is observed (US voter files). We contribute to many empirical social science literatures by providing a comprehensive comparative analysis of race imputation methods. By evaluating many available methods side by side, we provide researchers and policymakers with a practical framework for understanding and implementing race imputation techniques. This research aims to make race imputation methodology more transparent in social science, a goal that is increasingly relevant in an era where the proliferation of individual-level data has expanded the possibilities for demographic inference.

When evaluating performance, we describe three distinct objects of interest to researchers and practitioners: individual classification success, success at estimating racial disparities, and success at estimating racial demographics. Most comparisons of race imputation methods focus exclusively on the first object – individual classification success – by comparing statistics based on recall and precision rates. Therefore, another contribution of this paper is that three dimensions of performance are discussed concurrently.

Race can be imputed in simple or complex ways. It can be imputed from a single variable like name, non-text data like images, or the observation of online behaviors.[2] It can be based on simple probability rules based on population statistics, or can be the result of black-box machine learning methods trained on large labeled datasets. In this paper, to compare performance across methods, we limit ourselves to methods that we can replicate using publicly accessible data that reports race – namely, US voter file data. In practice, this means we compare imputation methods that require data on only names, only geography, or both names and geography in the US context.

This paper brings together race imputation methods that have have largely remained siloed within specific disciplines. While some techniques, like convolutional neural networks, are primarily utilized in computer science, others such as NamePrism have been confined to economics research. By bridging these disciplinary divides, we present what we believe to be one of the most comprehensive evaluation of race imputation methods to date.[3]

Our primary finding is that the best-performing race imputation method depends on the object of interest. In other words, no method we consider dominates across all types of performance. This is important since it means that practitioners and researchers must

[2]Race can be imputed from LinkedIn images (Berry et al., 2024). Facebook previously used onsite activity data to assign people to "ethnic-affinity groups," which Matthew et al. (2024) found were proxies for racial and ethnic group identification.

[3]Of course, this sort of performance evaluation is a moving target since new methods will keep being introduced. However, the goal of this paper is to provide an outline for current methods across fields and provide a framework for thinking through what good performance means in the context of race imputation.

consider their end goal when selecting race imputation methods.

Roadmap: Our paper proceeds as follows. Section 2 describes the related literature on race imputation. Section 3 outlines the different race imputation methods we compare in this paper. Section 4 explains the empirical strategy for testing performance across race imputation methods. Section 5 demonstrates performance across different methods. Section 6 discusses and concludes.

2 Related Literature

The closest related papers to ours are Kallus et al. (2022) and Decter-Frain (2022). Kallus et al. (2022) compares different proxies for race in mortgage lending (HDMA) data and medical data on drug dosing. Unlike in our context, the authors do not observe self-reported race – instead, they compare simulated results under different imputations. Decter-Frain (2022) benchmarks the most popular standard race imputation method (BISG) relative to custom ML imputations trained on voter registration files – he compares methods in terms of individual classification and estimating demographics (two of the three dimensions we discuss). He finds custom ML imputations outperform standard BISG in terms of individual classification but perform similarly at estimating demographics.

Administrative records of race may also differ from self-reported measures of race. Recent papers document widespread misclassification in administrative data and outline the consequent impacts on estimates of racial disparities. Baron et al. (2024) find that child protective service (CPS) investigators regularly misclassify the race of investigated children: around 9% of self-reported white children and 11% of self-reported Black children have a different race in the administrative data. The authors find that racial misclassification obscures around 24% of unwarranted racial disparity in foster care placement decisions. Finlay et al. (2024) find that 17% of defendants and 10% of prison inmates are misclassified in administrative criminal justice system records. This misclassification had clear consequences for descriptive statistics: the authors find the American Indian/Alaskan Native incarceration rate since 2010 was underestimated by 46% as a result. Jarrín et al. (2020) find that race in Medicare administrative data may differ from self-reports of race that are collected during routine home health care visits. Using race data from the Medicare enrollment database (EDB) underestimates the number of Hispanic people with dementia by 55% and the number of Hispanic people with diabetes by 59%.

Why might there be misclassification of race in administrative records? A few recent papers offer explanations. Finlay et al. (2024) outlines how operational and bureaucratic

complications make it difficult to correct misreports of race in the justice system. Luh (2022) studies how highway troopers record the race of motorists and finds that troopers may intentionally misreport motorist race to hide evidence of racial bias. Dobre and Witzen (2023) studies mortgage lender misclassification of race and finds that misclassification is more frequent when lenders are from a different racial group than the applicant.

This last example brings up a fundamental question in imputing race: which race do we want to impute? The one someone identifies with, or the one someone else perceives them to be? For the sake of studying discrimination, the third-party reported race might be more relevant than the self-report, an idea discussed implicitly by Wickett (2021), who studies colorism in the justice system using a prediction algorithm of race by facial features. This fundamental question is especially salient when it comes to more advanced imputation methods that integrate image-based information to impute race. For instance, both Berry et al. (2024) and Greenwald et al. (2023) study how to use images from LinkedIn to impute peoples' races. Imputed race from images might deviate from self-reports of race but align with a third-party perception of race.

Researchers may avoid imputation altogether if they can merge their data with other administrative data that has race recorded. For example, Ganong et al. (2023) link Chase bank account records to voter registration records that include self-reports of race and ethnicity, Bell et al. (2019) link patent records and tax records to NYC public schools data that includes records on race and ethnicity, and Choukhmane et al. (2023) links tax data to ACS data which contains race self-reports. However, linkage is not always feasible because of administrative costs or privacy concerns. King et al. (2023) describe in detail the privacy-bias trade-off in the context of US federal agencies. Laws protecting confidentiality can limit the capacity for data sharing, which leads some agencies (such as the US Treasury) to invest in imputation methods instead (King et al., 2023).

3 Outline of Different Race Imputation Methods

There are many ways to impute race. To impute, researchers must have access to other variables to form predictions. In this paper, to compare performance across methods, we limit ourselves to methods that we can replicate using publicly accessible data that reports race in the US. To the best of our knowledge, the only data that fits these requirements is US voter file data. In practice, this then means we compare imputation methods that require data on only names, only geography, or both names and geography.

We reviewed dozens of papers in statistics, economics, public health, and other fields that

use or study race imputation to get a collection of race imputation methods that we can take to the US voter files. Below are the methods that we outline and compare in the main text of the paper:[4]

1. **Last name:** imputes based on the most common race for a surname (based on census data).
2. **Zip:** imputes based on the most common race in the zip code (based on census data).
3. **Bayesian Improved Surname Geocoding (BISG)** imputes based on name and place (based on census data). This is the most common standard method for race imputation.
4. **Zest Race Prediction (ZRP)**: Matthews et al. (2024a) provides an open-source machine learning method to impute race based on full name and address. It is trained with data from voter file records linked to attributes from the 5-year ACS.
5. **NamePrism:** Ye et al. (2017) uses a probabilistic model based on historical data and patterns in names to impute race from names. It was trained with proprietary data from an email company.
6. **BIRDiE**: McCartan et al. (2023) is an imputation adjustment method. It estimates racial disparities by using surnames as a high-dimensional instrumental variable for race. It can be used in combination with any probabilistic imputation method.

When practitioners are limited in terms of available data, they are often limited in their available methods. If they have data on both name and place, they usually use BISG, which is the most popular race imputation method by far. In the following subsections, we give additional background on each of these methods.

3.1 Last name only

A very simple way of imputing race is to just predict based on last name only. There is 2010 census data on the number of people of different races with different last names. One can use that data to impute race as the most common race for each name. This method is commonly implemented by using the *predictrace* R package (Kaplan, 2023).

[4]We discuss more methods in a variety of Appendices. Appendix A.2 discusses race imputation methods that we do not compare due to the blackbox nature of their methods. Future appendices will compare results at different levels of geography, compare ZRP to other recent ML methods, compare results across thresholding options, and discuss results that require more input variables.

3.2 Geography-Only Methods

A straightforward but common way of imputing race is to predict race only based on location. Practitioners use ACS data on demographics by geographic unit for this procedure. For instance, one can impute race by setting race as the most common race within each county, tract, or zip code. In this iteration of the draft, we only discuss imputation by zip but in future drafts, we will compare this to county and tract approaches too.

3.3 Bayesian Improved Surname Geocoding (BISG)

The goal of the Bayesian Improved Surname Geocoding (BISG) methodology is to predict the race of an individual given an observed surname and geolocation. 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 method is implemented as follows.

The prior probability of a person's race is based on the Census files for the surname. A person with surname $j = 1, \ldots, J$ on the list and race, $i = 1, \ldots, 6$ has a prior probability given by $p(i|j)$, which is the proportion of all people with surname j who report being of race i in the surname file (the probability of a selected race given surname).

This probability is then updated based on the data from census on residences. Let us consider census zip code residence. For zip group $k = 1, ..., K$ then $r(k|i)$ is the proportion of all people in the Census file who self-report being race i who reside in Census zip k (the probability of a selected zip group of residence given race/ethnicity).

Then, $q(i|j,k)$ is the updated (posterior) probability of being of race/ethnicity i given surname j and census zip group of residence k. It is calculated using Bayes' rule as follows:

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

In a later development, Imai and Khanna (2016) developed an algorithm that used BISG as the baseline and improved upon it. The authors use Bayes's rule to combine the Census Bureau's Surname List with information in geocoded voter registration records (sourced from six states in the South that collect self-reported race information).[5] This has been

[5] Rosenman et al. (2023) describes the new dataset, which "expands on the available data for estimating race-name distributions." The state voter files used are those from Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina.

developed in a R package called *wru: Who are You? Bayesian Prediction of Racial Category Using Surname, First Name, Middle Name, and Geolocation*.[6]

In a recent paper, Imai et al. (2022) corrected for missing surnames in the Census files from which BISG draws. They supplement the census surname list with the lists of additional surnames, middle names, and first names from voter files. This innovation (called Bayesian Improved First Surname Geocoding (BISFG)) allows the authors to markedly reduce the proportion of missing surnames and further improve the race prediction. This has been already included in the updated version of the *wru* package in R.[7]

Custom BISG imputations: Note that BISG has been adapted to use additional variables internally by several organizations. For instance, CMS developed MBISG (Medicare BISG). The MBISG probabilities are based on the standard US Census data used for BISG, but also CMS's race administrative data and additional CMS data including first name, demographics, and coverage characteristics (Haas et al., 2019).

3.4 Zest Race Prediction (ZRP)

Zest Race Prediction (ZRP) was developed to be a BISG replacement by the company Zest AI (Matthews et al., 2024a).[8] ZRP imputes race using first, middle, last names as well as richer location data than used in standard BISG. ZRP advances beyond BISG by leveraging more input data (primarily richer location data) and by leveraging newer machine-learning methods to apply to the richer dataset (gradient boosting).[9]

ZRP models were trained using the 2021 voter registration files from Florida, Georgia, and North Carolina (Matthews et al., 2024b). The authors use voter registration data because it is "a large, publicly-available database of names, addresses and ground truth labels (self-reported race and ethnicity)" (Matthews et al., 2024b).[10] ZRP geocodes addresses

[6]The authors also modified the assumption from the BISG to include demographic variables (i.e. age and gender). This assumption states that given a voter's race, his/her surname does not contain any information about his/her geolocation and demographics. Letting x represent the two demographic variables age and gender, we can rewrite Bayes' rules as: $q(i|j,k,x) = \frac{p(i,x|j)r(k|i)}{\sum_{s=1}^{6} p(s,x|j)r(k|s)}$

[7]Clark et al. (2022) wrote an extension of the *wru* R package in which they combine ZIP codes and last names using Bayes' rule to implement BISG when zip code data is the geographic unit of interest.

[8]Zest AI started developing ZRP in 2020. The current version of ZRP (dated Dec 2024 as of Feb 2025) is open-source and available as a repository on Github Matthews et al. (2024a). ZRP has not been widely used at this point – the only academic paper we found that uses ZRP currently is Curiel and DeLuca (2024).

[9]Curiel and DeLuca (2024) describes ZRP as ML-modified BISG – the ML-based benefits come from how ML can "identify characteristics at the individual character (i.e., letter) level of surname, first, and middle names to improve upon the static [name dictionaries] employed by traditional BISG estimation."

[10]This is the same reason we and many other researchers use voter registration files for testing and validation for race imputation. The authors also note that "[i]deally a more comprehensive list of names

to their census block group, tract, or ZIP and then uses this to look up 5-year American Community Survey (ACS) attributes for the unit of geography. ZRP defaults to the smaller unit of geography possible from geocoding. ZRP methods use more recent ACS data than in standard BISG (which calls from 2010 census data).

The ZRP developers have shown evidence that ZRP outperforms BISG in a few ways. First, they show ZRP outperforms BISG in terms of classification accuracy using a subset of 230,000 Alabama voter file records (Matthews et al., 2024c). Second, they show ZRP outperforms BISG in terms of counts and classification accuracy for a nationwide dataset of 100,173 small business owners (Matthews et al., 2024c). Political scientists Curiel and DeLuca (2024) have also shown evidence of ZRP's benefits relative to BISG. They find that ZRP offers gains over BISG in estimating race for the purpose of redistricting, with the biggest gains being in racially diverse geographies where the Bayesian prior is not informative.

3.5 NamePrism

NamePrism trains a classifier by leveraging large private datasets and homophily in communications (Ye et al., 2017). (Similar people tend to associate with similar people.) NamePrism uses full name and 68 million full name and nationality pairs from a major email company and 6 million such pairs from Twitter for training their algorithm.

For common enough names, NamePrism uses data from the Census on race and last names, as other methods do. For first names, they use their email list of names that includes first and last names to estimate race distributions for first names based on the census distributions for corresponding last names. For less popular names, NamePrism relies on email and Twitter data.

NamePrism frames its contributions in a few ways. For one, it goes beyond previous classifier methods that used "small and artificial training sets, such as celebrity names from Wikipedia." Second, it moves beyond relying on substring features, instead using the email contact lists to derive "name embeddings" for classification. The authors show that their classification performs better than competing systems like *Ethnea*, *HMM*, and *EthnicSeer* in terms of individual classification success (Ye et al., 2017).

addresses and self-reported race/ethnicity from the US Census Bureau would be used to train the model, but such a dataset is not publicly available." (Matthews et al., 2024b).

3.6 BIRDiE

The final method we discuss is not so much a new imputation method, as it is an *imputation adjustment* method. BISG tends to produce well-calibrated racial predictions, but its residuals are often correlated with the outcomes of interest, yielding biased estimates of racial disparities. McCartan et al. (2023) propose an alternative identification strategy that corrects this bias. They call their method Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), which estimates racial disparities by using surnames as a high-dimensional instrumental variable for race.

Specifically, McCartan et al. (2023) assume that the outcome is conditionally independent of the surname given (unobserved) individual's race, residence location, and other observed attributes. This assumption is a type of exclusion restriction where surname serves as an instrumental variable for unobserved race. (It implies that for two individuals who live in the same area, belong to the same racial group, and share the observable attributes, their surnames have no predictive power of the outcome.)

Therefore, the main goal of BIRDiE is not to improve each individual-level prediction but to improve the measurement of some racial disparity of interest. BIRDiE can be used with any probabilistic imputation of race (not only BISG). In this paper, we implement BIRDiE corrections off of the baseline of BISG.

4 Empirical Strategy for Testing Performance across Race Imputation Methods

Data: To test performance, we impute race using datasets where the "ground truth" is known (i.e., where race is observed in the data). This way we can compare imputed race to observed race in the data. To the best of our knowledge, the only public US-based datasets that include name, place, and self-reported race are US voter files.

Six US states collect data about voters' self-reported race: Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina. These six states collect this data as a consequence of the 1965 Voting Rights Act – self-reported race data was collected to counter state-level voter disenfranchisement (McCartan et al., 2023; Cruz and Hayes, 2009). Collecting self-reported race was mandatory in these states for many years but became optional after the Supreme Court's Shelby County vs. Holder decision, which gutted the Voting Rights Act. Previously, the "[VRA] placed the burden of proof on government officials to prove why the changes they were seeking were not discriminatory. Now, voters

who are discriminated against now bear the burden of proving they are disenfranchised" (Levine and Rao, 2020).

For our empirical exercise, we use data from the 2023 North Carolina voter file. We use North Carolina because it is free and easy to access, which makes our replication exercise the most accessible to other researchers. 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 sample. In future drafts, we will conduct robustness exercises using voter files from other states too since Decter-Frain (2022) shows that imputation performance varies meaningfully by place.

Defining Performance: What does good performance mean when it comes to race imputation? There are several ways to define performance. For one, we may be interested individual-level classification performance. Performing well in this case means high precision and high recall (i.e., low numbers of false negatives and false positives). Second, we may be interested in performing well at estimating racial disparities of interest (i.e., describing differences across groups in the population accurately). Third, we may be interested in performing well at estimating population demographics (i.e., describing the population accurately).

Different methods may perform better on different dimensions. If an imputation is perfect (every imputation of race matches with reported race), it will necessarily dominate on all three fronts. However, when no imputation method is perfect, it is not necessarily the case that the highest-performing method in terms of individual classification outperforms on the other dimensions (even though many studies only report individual classification performance). For instance, one method may have higher recall and precision than another but perform worse on estimating racial disparities if its imputation of race is biased in a way that is correlated with outcomes of interest.[11] Therefore, it is necessary to compare methods across all three (individual, disparity, composition) types of performance.

4.1 Individual classification performance

We test methods' individual-level classification performance. To do so, we consider a few standard measurements: recall, precision, and F1 statistics. Each method has a recall, precision, and F1 scores for each specific racial group, g:

1. Recall is the fraction of people of group g who are imputed as group g. It is also

[11] Relatedly, Lu et al. (2024) shows bias, not variance, is likely the predominant type of error in BISG when it comes to estimating racial disparities with imputation.

known as the true positive rate and is calculated as $Recall = \frac{TP}{TP+FN}$ where TP is true positives and FN is false negatives.

2. Precision is the fraction of people imputed as group g who are truly of group g. It is also known as the positive predictive value and is calculated as $Precision = \frac{TP}{TP+FP}$ where TP is true positives and FP is false positives.

3. F1 scores are a common measure used to rate the success of a classifier. F1 scores are the harmonic mean of the two prior metrics (recall and precision). The formula for calculating these scores is $F1 = \frac{2\times(precision\times recall)}{precision+recall}$

All of the previous statistics give us a sense of performance by each racial group. How does one aggregate performance into an overall measure? There are three methods of aggregating F1 scores into one global F1 score:

1. The first method, the micro average, calculates positive and negative values globally. The formula used to calculate the micro average is: $F1^{micro} = \frac{TP}{(TP+0.5*(FP+FN))}$. Micro-averaging, therefore, aggregates the counts of true positives, false positives, and false negatives across all classes and then calculates the performance metric based on the total counts. This micro-average method gives equal weight to each instance, regardless of the class label and the number of cases in the class.

2. The second method, macro takes the average of each class's F1 score (i.e. simple average). Therefore, the macro-average gives equal weight to each class, regardless of the number of instances.

3. The third method is a weighted average, where the weights are the weights of the population in the sample by different races.

4.2 Estimating racial disparities

We are limited in what racial gaps we can estimate based on the available variables in the voter files. One of the only outcomes that is present in the voter file is party registration. Therefore, we discuss performance in terms of estimating Democratic party registration gaps across racial groups.[12] We do this first for the Black-White gap, as previously studied by McCartan et al. (2023). Then, we compare methods performance across all possible gaps in the population: Black-White, Hispanic-White, Asian-White, Black-Hispanic, Black-Asian, and Hispanic-Asian.

[12]If we wanted to estimate more economically interesting outcome variables, we would need to use restricted-access datasets for this entire paper.

4.3 Estimating racial demographics

Race imputation has been used to tell descriptive stories on the racial composition of populations of interest. For instance, Hepburn et al. (2020) uses court records with names and locations to impute race and discuss the likelihood of eviction by race. Similarly, Edwards et al. (2019) uses name and location to impute missing race information to study the likelihood of being killed by police by race. How would the racial composition of the population vary by imputation method?

We can compare the summary statistics on the population by imputation method. For a simple statistic that summarizes overall performance across methods, we can calculate the Euclidean distance between the distribution (as a point in 5-dimensional space) defined by the imputation and the real distribution (as a point in 5-dimensional space).

5 Results: Performance across Race Imputation Methods

The Voter File: We test performance using the North Carolina Voter File. This is publicly available data, which makes it a popular test case for race imputation papers. In the NC voter file, the following racial groups are recorded: Asian, Black or African American, American Indian or Alaska Native, Two or more races, Other, Native Hawaiian or Pacific Islander, Undesignated, White, and Hispanic. The standard racial categories for imputation methods are White, Asian, Black, Hispanic, and Other.

To map the racial 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. (In this data, then, 10% of observations fall into the Other category.) We discuss performance by input categories since the inputs available dictate the imputation methods available to the researcher.

Note on sample: We can impute race quickly using all methods in the main text except for NamePrism. NamePrism is available as a free web service for researchers.[13] However, the API limits the request rate to 60 requests per minute. Therefore, 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. We ran NamePrism in small batches over the course of a month to get predictions for the 500,000 observation sample. NamePrism, therefore, is a tool better suited for smaller samples than larger

[13] See API here: https://www.name-prism.com/api

samples.

5.1 Individual classification performance

Figure 1 summarizes recall and precision using F1 scores across racial groups. As a reminder, the F1 statistic is the harmonic mean of recall and precision.[14] There is an F1 score for each separate racial group in Figure 1. In a multi-class problem like race imputation, there are a few ways to aggregate the F1 statistics into one summary statistic. Figure 2 demonstrates the three distinct ways of aggregating group-specific F1 scores into one global F1 score and compares them across imputation methods.

The methods detailed include the three standard and best-known methods: name-only, geography-only, and BISG. The red bars show performance rates for using last name to impute, the pink bars show rates for using zip code only to impute, and the yellow bars show rates for using name and geography with BISG to impute. The two additional methods that are newer to the race imputation space are Zest Race Prediction (ZRP), shown in blue bars, and NamePrism, shown in green bars.

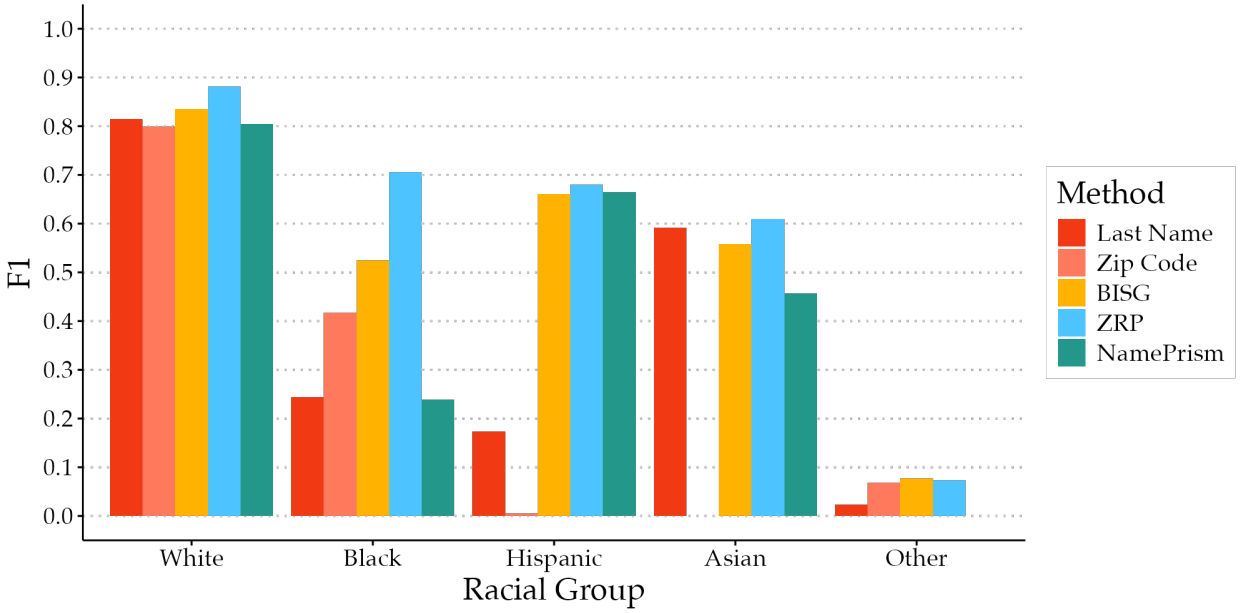
Figures 1 and 2 demonstrate a few known concepts among the standard methods. First, methods may perform better for some racial groups but not others. For instance, Figure 1 shows that using geography to impute race performs better than surname for predicting if someone identifies as Black. However, using geography gives F1 scores of near zero for predicting if people identify as Hispanic or Asian. Therefore, these figures replicate the known fact that geography-based imputation performs better for inferring Black identity, while surname-based imputation performs better for inferring Hispanic or Asian identity.[15]

Second, Figure 1 shows that the standard BISG methods improves predictions for most groups. The performance gains in moving from name or geography to the combination of the two are largest for Hispanic people and then Black people. However, BISG actually underperforms relative to surname for Asian people. In the aggregate, Figure 2 demonstrates that regardless of weighting method, BISG has a higher F1 statistic than the surname or geography methods. Third, Figure 1 shows that standard methods (last name, zip, BISG) all perform poorly for identifying people who are put in the Other group.

[14] Figures A.1 and A.2 show recall and precision rates by racial groups and imputations separately.

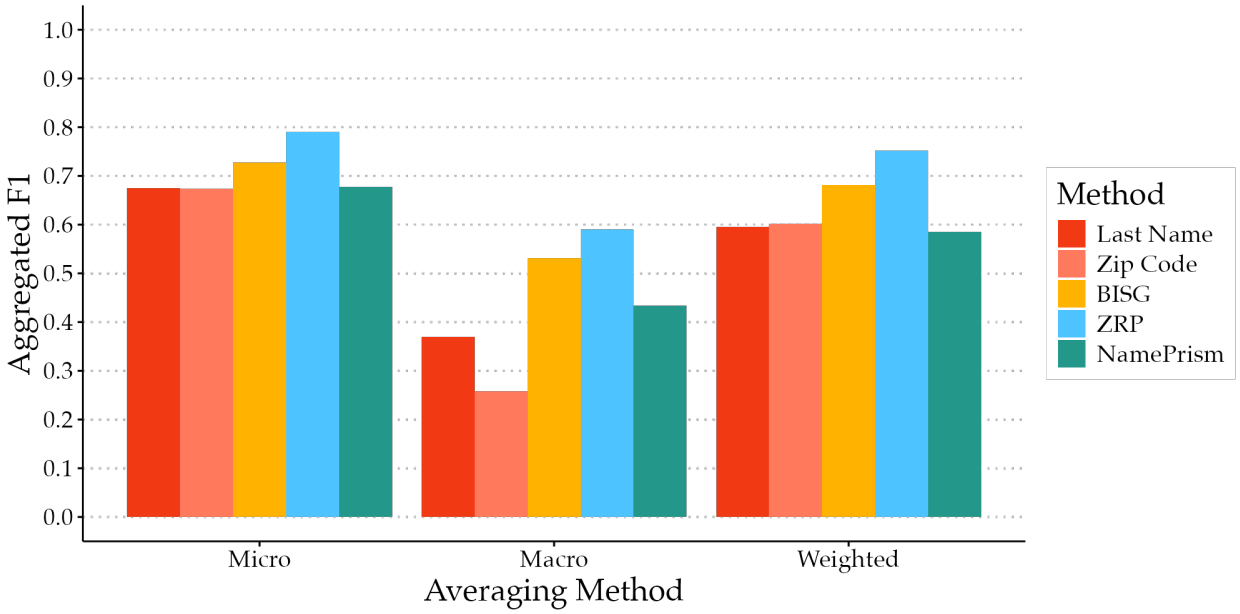
[15] This finding has been demonstrated previously by researchers such as Elliott et al. (2009) and Fiscella and Fremont (2006). This result is a consequence of the nature of geographic segregation and naming patterns in the US. Black-White segregation levels are higher than Asian-White or Hispanic-White 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 1: F1 Scores by Racial Group and Race Imputation Method



Notes: This figure shows the F1 scores across racial groups and race imputation methods using a 500,000 random sample of the NC voter file.

Figure 2: Aggregated F1 Scores by Race Imputation Method



Notes: This figure shows the aggregated F1 scores across race imputation methods using a 500,000 random sample of the NC voter file..

The results in Figures 1 and 2 go beyond previous work by comparing standard methods with ZRP and NamePrism. Figure 1 shows that ZRP dominates other methods across all racial groups with especially striking improvements for the Black population. 2 shows that regardless of weighting method, ZRP gives the best F1 score. Meanwhile, NamePrism underperforms many other methods overall and is especially poor performing for the Black population.

5.2 Estimating racial disparities

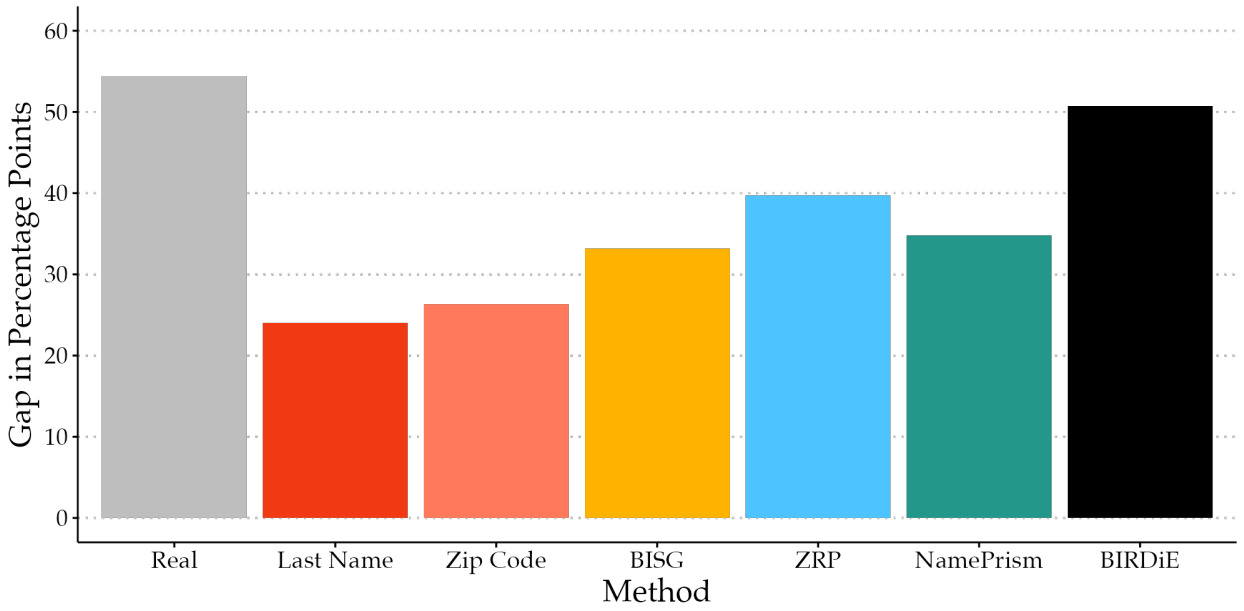
How do methods compare in terms of estimating racial disparities? First, we discuss the Black-White disparity in Democratic registration in the NC voter file. This is a nice example to test since it's simple to understand and it's observable in the NC voter file we use. It has also been previously studied by the author team who developed BIRDiE (McCartan et al., 2023).

Figure 3 demonstrates how the different methods perform when estimating the Black-white gap in Democratic registration in the NC voter file. In the data, 76% of Black registered voters are Democrats while 21% White registered voters are Democrats. Therefore, the true gap is about a 55 percentage point gap and is shown in gray. Figure 3 demonstrates that the standard methods (last name, zip, BISG) underestimate the true graph by 20-30 percentage points. ML-based methods improve on BISG to different degrees: NamePrism is a improvement over BISG by a few percentage points, while ZRP get the estimated gap up to 40 percentage points. ZRP, therefore, comes to closest to the true gap for the direct imputation methods.

Importantly, these results show that methods can do better in some dimensions of performance when they do worse in others. NamePrism outperforms BISG in estimating the Black-White racial gap (Figure 3), but NamePrism was markedly below BISG in terms of all the weighted F1 measures (Figure 2). Therefore, it is not necessarily the case that the method with the best F1 score will also have the most accurate measure of a given racial disparity.

The black bar in Figure 3 shows the estimated disparity when we use the BIRDiE adjustment. We replicate McCartan et al. (2023)'s finding that BIRDiE makes huge improvements in estimating the disparity of interest. The estimated gap is 50 percentage points, so it gets within 5 percentage points of the correct rate, which is an enormous improvement considering the standard BISG estimate is about 33 percentage points. Even though ZRP improves racial disparity estimation over BISG, it still does not improve estimation as

Figure 3: Black-White Disparity in Democratic Party Registration by Race Imputation Method



Notes: This figure shows the estimated Black-White disparity in democratic party registration by different race imputation methods. The gray bar is the real gap.

much as using the BIRDiE adjustment does.

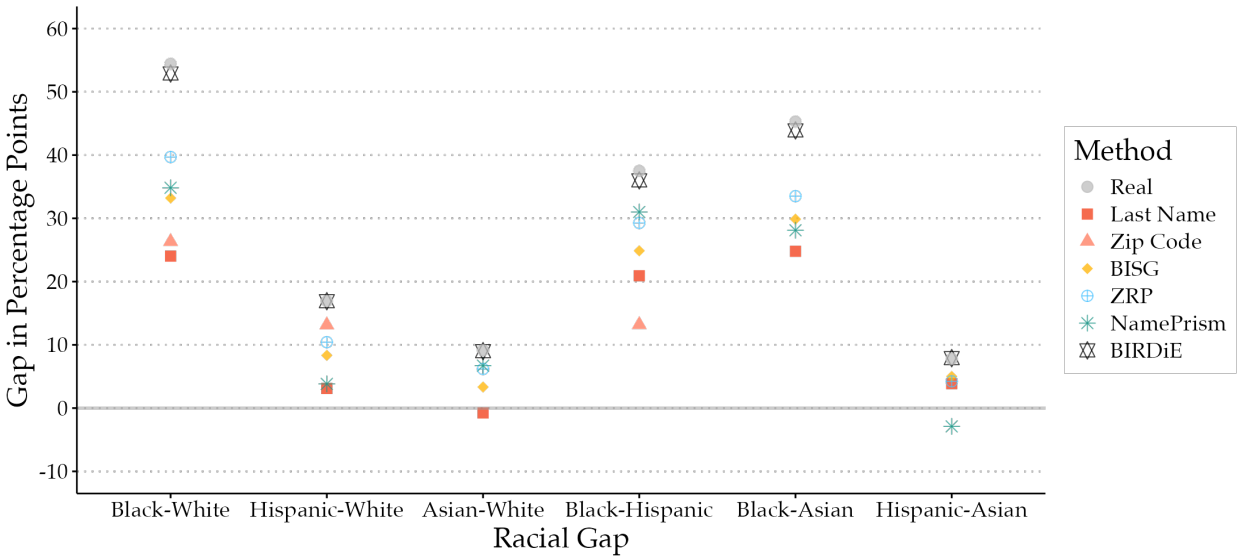
While Figure 3 demonstrates the Black-White gap across methods, this is not the only gap we care about. While the Black and White populations are the largest in this case, we also want to test performance across other gaps of interest. Therefore, we estimate six distinct gaps across all the methods: Black-White, Black-Hispanic, Black-Asian, Hispanic-White, Hispanic-Asian, and Asian-White. Figure 4 shows the results.

Figure 4 provides a number of insights. First, BIRDiE manages to get extremely close to the observed gaps for all combination of racial groups. Second, there are some cases where imputation methods might result in a gap that is the wrong direction altogether (see, for instance, the usage of NamePrism in estimating the Hispanic-Asian gap or the usage of last name in estimating the Asian-White gap). Third, the relative performance of some methods changes depending on which gap is of interest. For example, zip code provides the biggest underestimate of the Black-Hispanic gap, while it does the best out of the direct imputation methods for Hispanic-White. Put another way, the relative ranking of the methods is not consistent across the many possible racial gaps.

Overall, for practitioners, the evidence in Figure 4 highlights the importance of imputation method for estimating racial disparities. Gaps can reverse in direction based on choice of

imputation method, they can be even 30 percentage points away from observed gaps in the data, and their performance depends on the racial groups of interest. That said, Figure 4 is consistent in bolstering the existing case for using BIRDiE adjustment (McCartan et al., 2023).

Figure 4: 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. The plot shows the simulated gaps for six different gaps and across six different imputation methods. The gray dots are the observed gaps in the data.

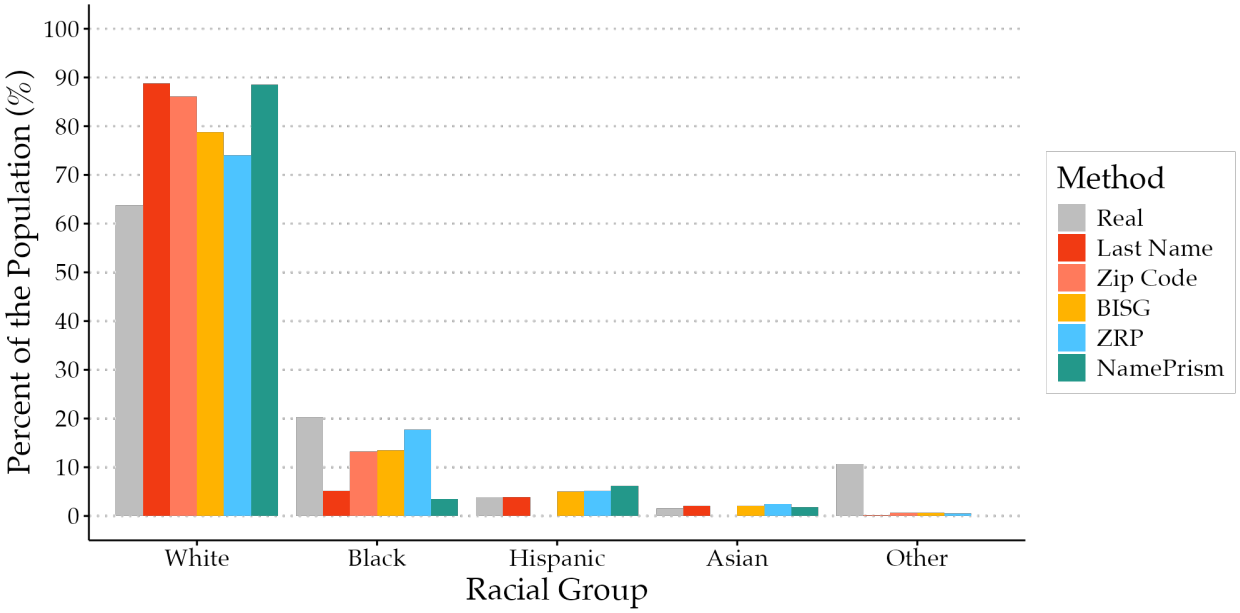
5.3 Estimating racial demographics

Finally, how do methods compare in terms of summarizing the populations accurately? I plot the composition by racial group and race imputation method in the NC voter file in Figure 5. The gray bars represent the actual population demographics, and the others represent the demographics under different imputation methods.

Figure 5 makes it clear that all methods significantly overestimate the percent of white people in the population, and all methods underestimate the fraction of Black people. No method does a good job of identifying people who do not identify with one of the four main racial groups (the true rate is about 10% in this voter file).

To provide a useful summary statistic on multi-dimensional performance, we calculate the Euclidean distance between the distribution (as a point in 5-dimensional space) defined by the imputation and the real distribution (as a point in 5-dimensional space). Table 1 shows that ZRP comes the closest to simulating the true population distribution (distance of

Figure 5: Racial Composition by Race Imputation Method (Name Methods)



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

14.74). The next best is BISG with a distance of 19.31.[16] Meanwhile, the ML-based method NamePrism is actually the furthest away from the true demographic distribution.

Table 1: Euclidean Distance between Simulated and Real Demographics across Race Imputation Methods

Method	White	Hispanic	Black	Asian	Other	Euclidean Distance
Real	63.75	3.75	20.29	1.58	10.64	NA
Last Name	88.76	3.83	5.20	2.04	0.18	31.03
Zip	86.05	0.07	13.25	0.00	0.62	25.75
BISG	78.73	5.06	13.46	2.08	0.66	19.31
ZRP	74.05	5.18	17.70	2.49	0.57	14.74
NamePrism	88.57	6.14	3.43	1.85	0.00	31.92

Notes: This table shows the simulated racial demographics by race imputation method in the NC voter file data. It also shows in the last column the euclidean distance between that prediction of demographics and the true demographics observed in the data.

Once again, these results demonstrate that methods can do better in some dimensions of performance but do worse in others. For example, NamePrism does worse than the standard methods (last name, Zip, BISG) on describing the population distribution (Figure

[16]Figure 5 shows that the primary way BISG mischaracterizes the population in this case is by missing 10 percentage points of Other people, 7 percentage points of Black people, and over-counting about 15 percentage points of White people.

5), but does better than those same methods in terms of estimating the Black-White racial gap (Figure 3).

6 Discussion of Results and Conclusion

In terms of researcher take-aways, standard race imputation methods can perform quite poorly for performance dimensions of interest. They systemically overcount White people relative to other groups, and they can dramatically miss the mark on estimated racial gaps (and can even reverse their direction). Two new methods make notable progress here.

First, Zest Race Prediction (ZRP), developed by Matthews et al. (2024a), outperforms other direct imputation methods on individual-level classification (Figure 2) and describing racial demographics (Table 1). However, while ZRP outperforms other direct imputation methods for some racial gaps, this is not true for others (Figure 4). The second new method – BIRDiE adjustment, developed by McCartan et al. (2023) – is a correction method for estimating racial disparities. We find BIRDiE leads to very accurate measures of racial disparities, and does so consistently across different racial group combinations (Figure 4).

This paper provides straightforward implications for practitioners in the form of methods recommendations, but it also brings up a number of broader concepts for quantitative researchers to consider when imputing race. For instance, our results show that methods can do better in some dimensions of performance when they do worse in others. If a method outperforms in terms of individual-level classification (F1 scores), this does not promise better estimation of racial demographics or racial disparities. Therefore, broader validation of imputation methods is needed to highlight the underlying assumptions and issues with these approaches.

These issues are especially important in practice because researchers might not always be able to use the best-performing tool. If researchers are only provided with BISG predictions with de-identified data, they might not be able to implement BIRDiE. If researchers only have data on names, they cannot use methods that require geography data like BISG or ZRP. When researchers are limited in possible imputation methods, they should consider and discuss how the shortcomings of imputation methods feed into the measurement of their objects of interest, and their subsequent conclusions.

References

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*.

Bell, A., R. Chetty, X. Jaravel, N. Petkova, and J. Van Reenen (2019). Who Becomes an Inventor in America? The Importance of Exposure to Innovation. *The Quarterly Journal of Economics 134*(2), 647–713.

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.

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.

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.

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*.

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

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 (2024). 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*.

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.

Egan, M., G. Matvos, and A. Seru (2022). When Harry Fired Sally: The Double Standard in Punishing Misconduct. *Journal of Political Economy 130*(5), 1184–1248.

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.

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*.

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.

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

Greenwald, D., S. T. Howell, C. Li, and E. Yimfor (2023). Regulatory Arbitrage or Random Errors? Implications of Race Prediction Algorithms in Fair Lending Analysis.

Haas, A., M. N. Elliott, J. W. Dembosky, J. L. Adams, S. M. Wilson-Frederick, J. S. Mallett, S. Gaillot, S. C. Haffer, and A. M. Haviland (2019). Imputation of Race/Ethnicity to Enable Measurement of HEDIS Performance by Race/Ethnicity. *Health services research 54*(1), 13–23.

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.

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.

Kallus, N., X. Mao, and A. Zhou (2022). Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. *Management Science 68*(3), 1959–1981.

Kaplan, J. (2023). predictrace: Predict the Race and Gender of a Given Name Using Census and Social Security Administration Data. Version 1.7.1. Available online at https://jacobkap.github.io/predictrace/.

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.

Levine, S. and A. Rao (2020). In 2013 the Supreme Court gutted voting rights—how has it changed the US? *The Guardian*. Accessed: 2024-10-02.

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

Lu, B., J. Wan, D. Ouyang, J. Goldin, and D. E. Ho (2024). Quantifying the Uncertainty of Imputed Demographic Disparity Estimates: The Dual-Bootstrap. Technical report, National Bureau of Economic Research.

Luh, E. (2022). 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). Zest Race Predictor. https://github.com/zestai/zrp.

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

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

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.

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.

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.

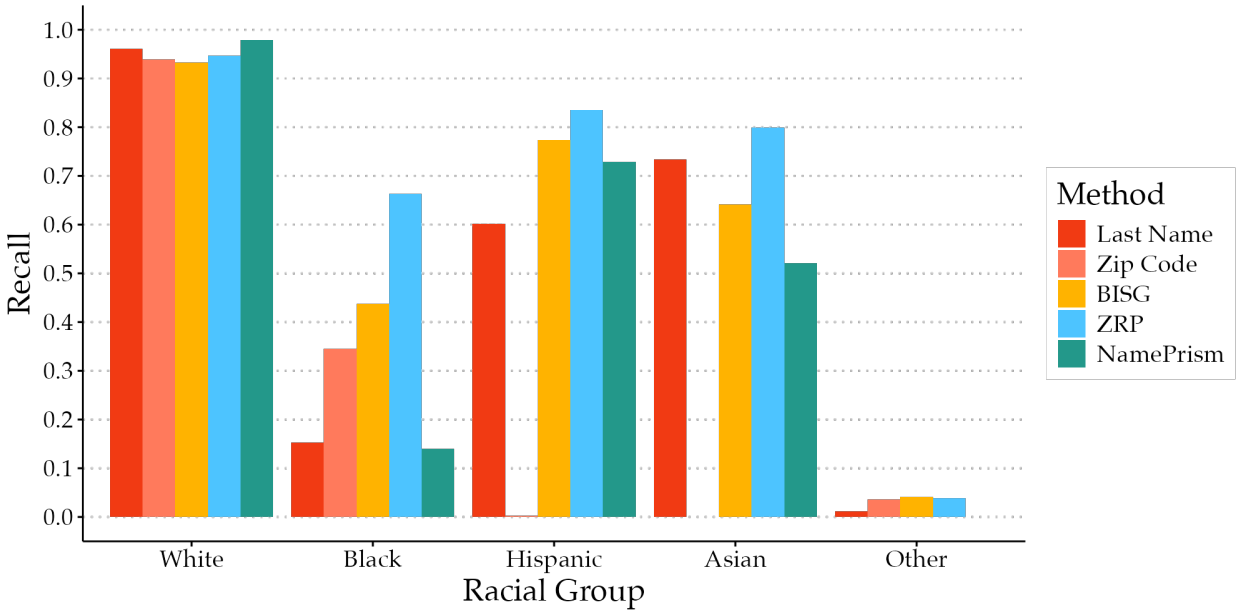
Wickett, A. (2021, June 28-July 2). Not so Black and White: An Algorithmic Approach to Detecting Colorism in Criminal Sentencing. In *Proceedings of the ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*. ACM.

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

Appendix

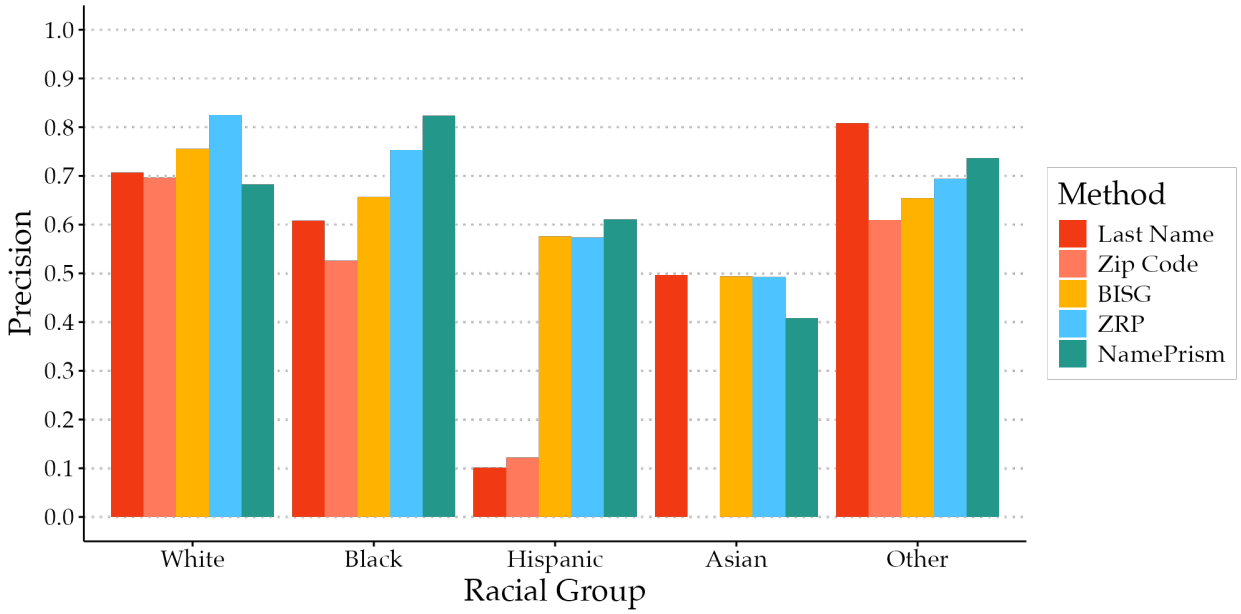
A.1 Supplementary Figures and Tables

Figure A.1: Recall Rates by Racial Group and Race Imputation Method



Notes: This figure shows the recall rates across racial groups and race imputation methods using a 500,000 random sample of the NC voter file.

Figure A.2: Precision Rates by Racial Group and Race Imputation Method



Notes: This figure shows the precision rates across racial groups and race imputation methods using a 500,000 random sample of the NC voter file.

A.2 Black Box Methods

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 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.

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

This paper, therefore, limits itself to comparing free methods for race imputation with technical documentation. In this draft, we do not extend the exercise to include black box tools for purchase.

Side note: NamSor has more in common with NamePrism than other methods. That is because both have use cases outside of the US context. NamSor and NamePrism are unique in that they have international use cases.