Letter

Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records

LISA P. ARGYLE   *Brigham Young University, United States*

MICHAEL BARBER   *Brigham Young University, United States*

*W*e show that a common method of predicting individuals' race in administrative records, Bayesian Improved Surname Geocoding (BISG), produces misclassification errors that are strongly correlated with demographic and socioeconomic factors. In addition to the high error rates for some racial subgroups, the misclassification rates are correlated with the political and economic characteristics of a voter's neighborhood. Racial and ethnic minorities who live in wealthy, highly educated, and politically active areas are most likely to be misclassified as white by BISG. Inferences about the relationship between sociodemographic factors and political outcomes, like voting, are likely to be biased in models using BISG to infer race. We develop an improved method in which the BISG estimates are incorporated into a machine learning model that accounts for class imbalance and incorporates individual and neighborhood characteristics. Our model decreases the misclassification rates among non-white individuals, in some cases by as much as 50%.

INTRODUCTION

In recent years, scholars have developed methods for predicting the ethnicity of individuals using their surname and geographic information (Elliott et al. 2009; Hofstra and de Schipper 2018; Imai and Khanna 2016). At their core, all of these methods use Bayes' theorem to generate a probability that a person belongs to a particular race using the individual's surname and geographic location. These probabilities are estimated from information from the U.S. Census Bureau that indicates the relative frequency of surnames among different racial groups and the racial composition of neighborhoods. These Bayesian Improved Surname Geocoding (BISG) methods have been applied in a variety of different substantive areas where information about a person's race would be valuable and informative. For example, the Imai and Khanna (2016) method of predicting an individual's ethnicity from information contained in voter registration files has become widely used in the few years since its introduction. Scholars have used this method to compare across racial groups' political behavior (Enos, Kaufman, and Sands 2019; Fraga 2018; Grinberg et al. 2019; Grumbach and Sahn 2020), lending in housing and auto markets (Baines and Courchane 2014; Thomas 2017), criminal justice (Edwards, Esposito, and Lee 2018; Edwards, Lee, and Esposito 2019), health and medical outcomes (Adjaye-Gbewonyo et al. 2014; Nguyen et al. 2019), and policymaking (Einstein, Glick,

and Palmer 2019; Henninger, Meredith, and Morse 2018). Moreover, the method has been used in a variety of nonacademic situations such as election forecasting and state and federal lawsuits.[1]

Given the limited information available to the BISG model, many of the predictions of voters' ethnicities will be incorrect, something that previous research has noted (e.g., Elliott et al. [2009], Imai and Khanna [2016], and Voicu [2018], who extends these models to include first names in addition to surnames).

In this research note, we bring to light and quantify an important and previously unidentified problem with these racial classification models. While many other studies have shown the overall misclassification rates by race (Adjaye-Gbewonyo et al. 2014; Baines and Courchane 2014; Elliott et al. 2009; Fiscella and Fremont 2006; Martino et al. 2013), none have identified the correlation between misclassification and political and socioeconomic factors.[2] For example, among Blacks, the rate of misclassification is highly correlated with the income and socioeconomic status (SES) of the individual's census tract. Thus, Black individuals in wealthy neighborhoods are more likely to be misclassified by the model than are Black people in lower income neighborhoods. Insofar as income is correlated with factors that a researcher may be studying—such as political preferences, economic behavior, or health outcomes—any comparisons between individuals predicted to be Black versus white will suffer from significant, systematic bias.

We demonstrate the misclassification bias of Imai and Khanna's (2016) model using the Florida and

Lisa P. Argyle [iD], Assistant Professor, Department of Political Science, Brigham Young University, United States, lpargyle@byu.edu.
Michael Barber [iD], Associate Professor, Department of Political Science, Brigham Young University, mbarber@byu.edu.

---

[1] See, e.g., N.A. for Advancement of Colored People v. E. Ramapo Cent. Sch. Dist. No. 17-CV-8943 (CS) (S.D.N.Y. May. 25, 2020).
[2] Baines and Courchane (2014) is an exception and notes that the BISG model performs worse for African Americans and Hispanics as "FICO scores and incomes rise" (157).

North Carolina voter registration files. We then propose an ensemble machine learning approach to reducing the bias of misclassification error. This approach incorporates the BISG estimates into a second model with additional demographic information and an algorithm that prioritizes correct classification of minority groups. Our proposed refinement achieves dramatic reductions in the correlation between misclassification error and SES variables.

DATA AND RESULTS

We obtained the 2018 Florida and North Carolina voter registration files from the data and analytics firm The Data Trust, LLC. The Florida and North Carolina voter files contain the address and self-reported race of each registered voter in each state. We then combined this with demographic information for each voter's census tract.

To benchmark the model against the results in Imai and Khanna (2016), we begin by implementing the predicted race model on the 2018 Florida and North Carolina files and obtain the overall error rate by comparing the predicted ethnicity for each voter to the voter's self-reported race.[3] While our 2018 files are more recent, the overall results are similar to those reported in Table 1 of Imai and Khanna (2016). The model allows for the inclusion of additional variables beyond surname and census tract, including political party, age, and gender. We present the results of the model that uses surname, census tract, and party as that is the model used in Imai and Khanna (2016). However, the results of models that include additional covariates are shown in A.1 in the Supplementary Material. The substantive conclusions reached do not depend on the BISG specification.

The BISG model has a 15.1% error rate in Florida and a 15.4% error rate in North Carolina.[4] However, this rate varies dramatically across race and ethnicity. In Florida and North Carolina, 6.7% and 6.9% of self-identified white voters are incorrectly classified as being non-white. On the other hand, 24.7% and 31.6% of voters who self-identify as non-white are incorrectly predicted to be white.[5] Future researchers are advised that in some applications, using predicted probabilities may be more appropriate than a deterministic racial classification based on highest probability.

Looking further down the BISG columns shows that in some areas, the model excels and in others it struggles significantly. The model is very good at avoiding false positives among minority groups (i.e., incorrectly classifying people as belonging to a minority group). For example, in Florida (North Carolina), the false positive rate is only 2.9% (5.1%) for Blacks, 3.7% (1.5%) for Hispanics, and less than 1% (0.8%) for Asians. However, the model has very high false negative rates among minority groups (i.e., incorrectly classifying a Black person as not being Black). This can be seen in the very high rates of false *negatives* for Blacks (33.5% FL, 34.1% NC), Hispanics (15.1% FL, 23.7% NC), and Asians (47.5% FL, 34.0% NC). In other words, among self-reported Black voters, roughly one-third of them are incorrectly classified as non-Black. These false-negative and false-positive rates are similar to those shown in Imai and Khanna (2016).

Table 1 largely replicates results in Imai and Khanna (2016); however, an unanswered question that remains is whether the misclassification rate is correlated with other factors. If the misclassification rates in Table 1 occur more or less at random, this would be less concerning than if the classification model is systematically wrong for particular types of individuals. Systematic misclassification could introduce significant bias into any analyses that use predicted race.

Figure 1 shows four panels, one each for self-identified white, Black, Hispanic, and Asian individuals. In each panel, the x-axis is the median income of the individual's census tract. The y-axis is the proportion of individuals whose race is misclassified by the BISG model. Each point shows the binned (by $1,000 increments) misclassification rate, and the red line is a weighted (by population size) lowess curve fit to those binned points. The pattern of misclassification varies dramatically by race and tract income. For whites, misclassification rates are relatively low overall and are also negatively correlated with income—that is, individuals who live in the poorest neighborhoods are the most likely to be misclassified by the model. Among Black individuals, the trend is strikingly large and in the opposite direction. Among Blacks who live in the wealthiest census tracts, the misclassification rate is nearly 100%, meaning that the model is inaccurately classifying nearly all Blacks living in wealthy census tracts. Furthermore, the misclassification rate is quite high even in modestly wealthy census tracts. The average misclassification rate approaches 50% around census tracts with a median income between $50,000 and $60,000. Among Hispanics, there is a positive correlation between census tract income and model misclassification; however, the degree of change across the figure is not as large as among Blacks. Finally, among Asian voters, there is no linear trend, but rather an S-shaped relationship, and the overall misclassification rate remains very high (approximately 50%) across all levels of income.[6]

---

[3] We use the R package wru created by Imai and Khanna (2016)—version 0.1-9, available at https://cran.r-project.org/web/packages/wru/wru.pdf.

[4] The F1-score is a measure of model accuracy based on a combination of precision and recall. Higher values indicate better performance. Section A.4.3 of the Supplementary Material discusses this measure in more detail.

[5] We classify each voter's predicted race as the ethnicity over which the model places the highest probability. Alternative methods of classification yield similar results. See Figure A.7 in the Supplementary Material.

[6] Figure A.4 in the Supplementary Material shows these same figures, but for a BISG model that also includes party, gender, and age in addition to surname and census tract. Figure A.5 in the Supplementary Material shows that the bias persists across other measures of SES, including education, home ownership, vote propensity, and campaign contributions.

TABLE 1. Replication and Extension of Imai and Khanna's (2016) Race Classification Model using 2018 Florida and North Carolina Voter Files

		Florida		North Carolina	
		BISG: Surname Census Tract Party	Random forest	BISG: Surname Census Tract Party	Random forest
Overall error rate		0.151	0.142	0.154	0.148
F1-score		0.601	0.630	0.581	0.587
White (FL: 64.8%; NC: 71.2%)	False positive	0.247	0.189	0.316	0.255
	False negative	0.067	0.082	0.069	0.083
Black (FL: 13.8%; NC: 23.0%)	False positive	0.029	0.041	0.051	0.065
	False negative	0.335	0.231	0.341	0.271
Hispanic (FL: 16.8%; NC: 2.9%)	False positive	0.037	0.035	0.015	0.017
	False negative	0.151	0.141	0.237	0.226
Asian (FL: 2.0%; NC: 1.3%)	False positive	0.008	0.007	0.008	0.007
	False negative	0.475	0.476	0.340	0.348
Other (FL: 2.7%; NC: 1.6%)	False positive	0.000	0.004	0.002	0.001
	False negative	0.997	0.949	0.994	0.992

Note: Numbers in parentheses in the first column represent the proportion of voters in the voter file who self-identify with each racial category in FL and NC, respectively. The two columns labeled "BISG" show the classification error rates using the Imai and Khanna (2016) wru model. The "Random forest" columns show the results using our proposed improvement. The random forest model is trained on 80% of the FL data, and FL results in in this table are predictions on a 20% test set. Results for NC are a true out-of-sample prediction, based on the FL-trained model.

Whereas Figure 1 is plotting the false-negative rate (i.e., incorrectly predicting a Black person to be non-Black) across tract income, Figure A.3 in the Supplementary Material shows the false-positive rate (i.e., incorrectly classifying a non-Black person as Black) for each of the four racial groups. The results show that the high false-negative rate for non-white voters at high incomes (especially among Black voters) is due to a high false-positive rate in which the model incorrectly predicts a person to be white.

The intuition behind the correlation between socioeconomic factors and misclassification rates is quite simple: if a voter's surname is relatively common across multiple ethnicities (i.e., Brown or Johnson for white and Black individuals), then the model will lean more heavily on inferring race from the distribution of ethnicities in the individual's census tract. However, this will lead to errors for voters who are racially distinct from the majority of their neighbors—that is, "local minorities." This pattern is especially true among white and Black voters because of the relative similarity of surnames between whites and Blacks, which causes the model to rely more heavily on geographic factors when making its predictions. This is less the case among Hispanics and Asian individuals for whom surnames tend to be more ethnically distinct. Figure A.6 in the Supplementary Material shows that people who live in areas where they are local minorities are much more likely to be misclassified.

Another consideration is the degree to which the racial prediction model misclassifies people's race because of factors related to gender. The connection to gender is likely due to the fact that women are significantly more likely than men to change their surname after marriage. This, combined with the different propensities for interracial marriage across ethnicities, suggests that minority women are especially difficult to correctly classify. Figure A.10 in the Supplementary Material shows that this is the case.

IMPLICATIONS FOR RESEARCHERS AND PROPOSED IMPROVEMENT

The BISG method of predicting race from census and geographic data provides scholars a previously unavailable opportunity to incorporate racial considerations into important questions of political behavior, representation, accountability, and material well-being. However, this advancement comes with significant limitations. If racial misclassification is highly correlated with political and economic factors, estimation may be biased when incorporating these predictions into any study of political economy.

We propose an ensemble solution for improving the accuracy and minimizing misclassification biases of the BISG racial classification. We incorporate the Imai and Khanna (2016) BISG probability scores for each racial class as inputs into a random forest model with other individual- and neighborhood-level political and socioeconomic predictors. The intention is to use a second model to adjust the BISG predictions so that they become less correlated with political and economic factors. The random forest is a straightforward and widely used machine learning tool for multiclass prediction. In brief, the random forest identifies potential splits in the data that maximize the accuracy of classification of observations into each racial

FIGURE 1. Misclassification Rates and Census Tract Income
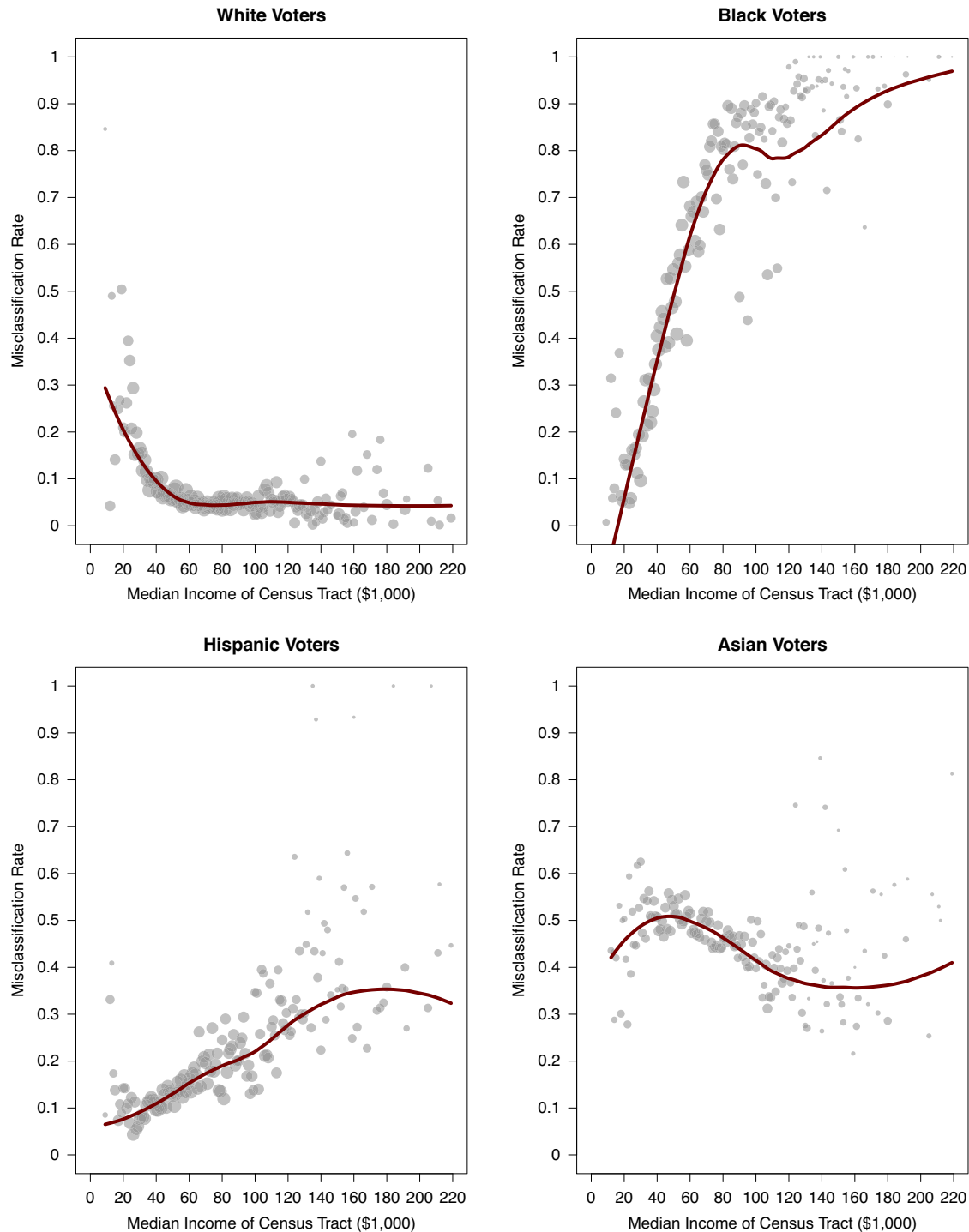
Note: Each panel shows the relationship between voters' census tract income (*x*-axis) and the proportion of voters from each race that are misclassified by the wru model, using surname and census tract. Points, sized in proportion to number of observations, show average misclassification rate for each $1,000 increment. The line plots a lowess fit (span = 0.6) through those points, weighted by number of cases.

subgroup. Because of class imbalance (the majority of Florida voters are white, whereas other racial groups, especially Asian and "Other," are much smaller), we incorporate class weights into the model and use a macro-averaged F1-score rather than total prediction error to select the best model parameters. Both of these changes place an additional modeling emphasis on correct classification of racial and ethnic minorities,

relative to correct classification of white voters. The model is trained and tuned using a randomly selected 80% subset of the Florida voter file and then "tested" on the remaining 20% of the voter file. Section A.4 of the Supplementary Material provides a complete description of training, tuning, and testing the random forest model.

We test the random forest approach to classification on two different datasets and show how it both improves classification and leads to more accurate inferences based on those predictions. First, we apply the parameters from the random forest model to the entire North Carolina voter file for a true out-of-sample prediction of each voter's race. Second, we generate out-of-sample prediction on the held-out 20% of the Florida voter file. We focus here on the North Carolina data and present the results from the 20% Florida sample in the Supplementary Material. Racial classification predictions are of most value in contexts where racial data are not readily available, but when applying a trained model from one state to another, the researcher must make the assumption that the relationship between a person's name, geographic location, political involvement, income, and other neighborhood characteristics and their racial identification is the same in both locations. Because no North Carolina data are used in training the random forest model, this provides a true out-of-sample test of the BISG + random forest model, which allows us to examine the performance of the predictions across state contexts given these assumptions. We find that the model performs well in both North Carolina and the withheld test data in Florida. Future researchers adopting this method should carefully consider whether this assumption continues to hold in their specific application.

While we test the accuracy of our solution using Florida and North Carolina, where self-reported race is available, we emphasize that the proposed solution, similar to the original BISG model, can be implemented anywhere that researchers have geographic location data and individual surnames, plus other desired covariates. To allow future researchers to implement our proposed improvements more easily, we have included the random forest model object in this article's replication materials (Argyle and Barber 2023). In this way, future researchers with data that include the same set of publicly available predictor variables can use it for racial classification predictions without needing to retrain their own random forest model, as we demonstrate with the North Carolina predictions in the article.[7] Alternatively, if researchers have applications in a different domain where any different set of predictor variables are available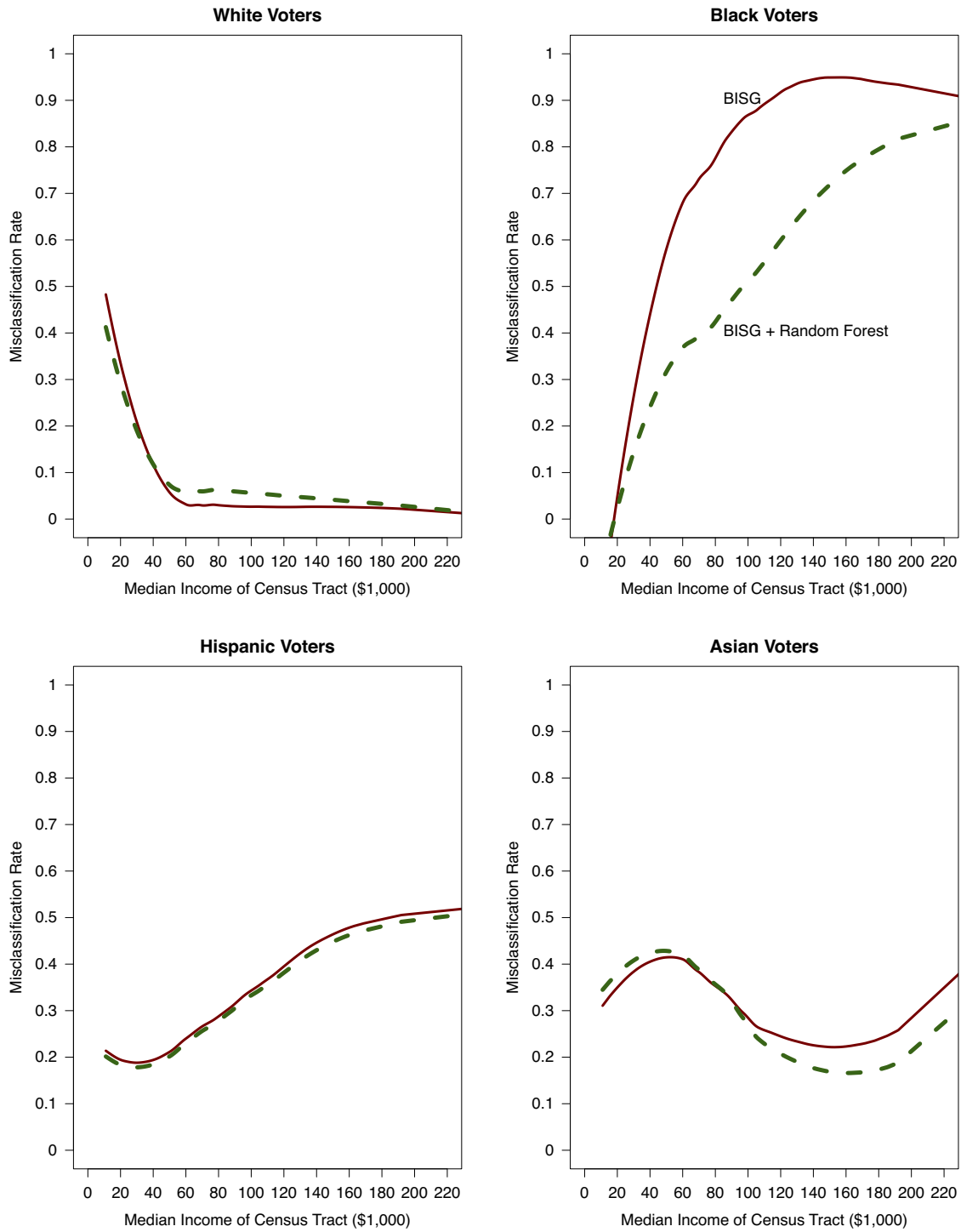 and theoretically important, they can adapt our public replication code to train their own random forest model. This will require a set of training data in which self-reported race is available, in addition to the data for which the researcher intends to make racial classification predictions. This code uses straightforward commands in existing R packages.

We calculate the overall misclassification rates as well as the false-positive and false-negative rates for each race using the random forest improvement. Because such a high proportion of voters in Florida and North Carolina are white, overall accuracy of the BISG model can look quite good if it selects "white" for all uncertain cases. The class weights and macro-F1 optimization reduce the propensity to select the major class in uncertain cases, so that the largest and most important gains are expected to be in a reduction of false positives among white respondents and a larger reduction in false negatives among minority populations. The cost of this approach can be a reduction in the accuracy of predictions of the major class (white) because some uncertain cases that were "guessed" correctly are no longer accurately classified.

The misclassification, false-positive, and false-negative rates are reported next to the rates for the original BISG model in Table 1. The results are as expected: the BISG + random forest algorithm generates lower overall misclassification rates compared with the BISG model across all individuals of all races, but the improvements are relatively small in magnitude (around 1 percentage point). However, more dramatic improvements come when looking at the false-positive and false-negative rates for non-white respondents. For example, the random forest model in North Carolina reduces the false-negative rate among Blacks from 0.341 to 0.271, a 20.53% reduction, and the false positive rate for whites declines from 0.316 to 0.255, a 19.3% reduction. The improvements are even larger in Florida. We note that, as expected, there is a small increase in the false-negative rate of white respondents in both states. Improvements are quite small among Hispanic, Asian, and "Other" racial groups, which suggests that the ensemble method is unlikely to come at a cost of reduced accuracy for any racial group.

The improvements are even larger in certain ranges of census tract income. Figure 2 shows a reduction in the correlation between misclassification and census tract income, particularly among Blacks. The solid red lines show the same misclassification rates as Figure 1 for the BISG model. The dotted green line in Figure 2 shows the misclassification rate across census tract income for the BISG + random forest algorithm. Among Black individuals, across all income ranges, the random forest model produces lower misclassification rates than the original BISG model. In some areas (around $80–$100 k tract median income), the misclassification rate is reduced by nearly 50%. Figures A.11 and A.12 show the same results in the withheld 20% of the Florida voter file.

---

[7] The variables in our model are listed in Section A.4.2 of the Supplementary Material. They are derived from public data sources, and include the BISG probabilities, individual political party, sex, and age, neighborhood socioeconomic, racial, and population data, and campaign donation histories.

FIGURE 2. Misclassification Rates and Census Tract Income with Random Forest Improvements

Note: Misclassification rates and census tract income for the BISG model and the BISG + random forest algorithm in North Carolina. The solid red line shows the average misclassification rate for the BISG model and the dotted green line shows the misclassification rate for the BISG + random forest algorithm using a lowess line (span = 0.6) fit to the data. The addition of the random forest algorithm dramatically decreases misclassification rates among Black voters.

Table 2 looks at several measures of economic and political inequity across races using self-reported race, the BISG model alone, and the BISG + random forest model. Each column of the table presents a different theoretically and empirically important variable from the full North Carolina voter file: income, home value,

TABLE 2.   Differences in Summary Statistics for Self-Reported Race versus Predicted Race Using BISG and BISG + Random Forest Models in North Carolina

|		Median income	Median home value	Campaign donors	2016 turnout percentage	Minority in own tract
White:	Self-reported	$62,002	$211,393	84.36%	62.92	5.73%
	BISG model	$62,273	$210,862	88.11%	69.05	4.20%
	% Diff.	0.44	−0.25	4.45	9.73	−26.66
	BISG + RF model	$62,450	$212,091	84.20%	64.08	4.14%
	% Diff.	0.72	0.33	−0.19	**1.84**	−27.72
Black:	Self-reported	$48,435	$164,471	12.32%	58.6	57.66%
	BISG model	$41,193	$146,829	8.79%	38.72	32.67%
	% Diff.	−14.95	−10.73	−28.64	−33.93	−43.34
	BISG + RF model	$46,435	$159,960	13.06%	55.52	50.64%
	% Diff.	**−4.13**	**−2.74**	**6.00**	**−5.27**	**−12.18**
Hispanic:	Self-reported	$57,197	$199,815	0.73%	29.87	100.00%
	BISG model	$56,255	$198,056	0.89%	38.08	100.00%
	% Diff.	−1.64	−0.88	21.67	27.47	0.0
	BISG + RF model	$56,035	$199,612	0.87%	39.67	100.00%
	% Diff.	−2.02	**−0.10**	**19.20**	32.8	0.0
Asian:	Self-reported	$75,814	$258,025	1.31%	35.66	100.00%
	BISG model	$78,475	$266,070	1.90%	42.76	100.00%
	% Diff.	3.51	3.12	44.94	19.91	0.0
	BISG + RF model	$81,495	$272,991	1.85%	40.77	100.00%
	% Diff.	7.49	5.80	**40.65**	**14.33**	0.0

Note: Among Black individuals, predicted race using the BISG model leads to estimates of lower median income, home value, rates of campaign donations, and rates of living as a minority in one's census tract. The BISG + random forest ensemble improves estimates. The "Campaign donors" column measures the estimated proportion of donors who identify with each ethnicity. The "Minority in own tract" column measures the proportion of individuals from that racial group who live in a tract in which their race is not the largest group. Bold values indicate cases where the random forest model provides improvements over the BISG model alone.

campaign donations, voter turnout, and geographic segregation.[8] In many cases, the differences in these factors using self-reported race and the BISG model are quite large, especially among Black individuals. The next row in the table shows these same estimates using our proposed BISG + random forest solution. Incorporating our proposed solution of using the BISG model with the random forest correction improves estimates (brings them closer to the estimates that use self-reported race) in nearly all cases. For example, the BISG model underestimates Black median tract income by 14.95% but has a much smaller underestimate of 4.13% using the BISG + random forest model. The same is true of median home value (−10.73% using BISG alone vs.−2.74% using BISG + random forest), the share of campaign donors who are Black (28.64% using BISG alone vs. 6.00% using BISG + random forest), voter turnout (−33.9% using BISG alone vs. −5.27% using BISG + random forest), and whether an individual is a minority in their own neighborhood

(i.e., lives in a census tract where their race is not the most common racial group,−43.3% using BISG alone vs.−12.18% using BISG + random forest). Bold values indicate cases where the random forest model provides improvements over the BISG model alone.

Finally, we look more deeply at how the improvements generated by the random forest model alter the substantive conclusions a person might draw regarding voter turnout compared with the conclusions they would arrive at based solely on the BISG model. Scholars have long discussed the variation in turnout by race and income (Fraga 2018; Wolfinger and Rosenstone 1980). In fact, a key component of the U.S. Supreme Court's decision to strike down portions of the Voting Rights Act in *Shelby County v. Holder* relied on estimates of turnout rates by race (Ansolabehere, Fraga, and Schaffner 2020). Figure 3 shows the estimated turnout rate among white and Black voters in North Carolina and the income of their census tract. Calculating turnout rates requires a numerator (the number of people who voted) and a denominator (the number of people who are eligible to vote). We rely on the voter file to provide the number of individuals by race who turned out to vote in the 2016 general election, as other scholarship has argued that voter files provide the most accurate measure of how many voters turned out in an election (Fraga and Holbein 2020). We also follow the scholarly convention of using the U.S. Census estimates of

---

[8] For example, see Herring and Henderson (2016) for BISG use in income/wealth measurements, Grumbach and Sahn (2020) for BISG use in campaign contributions, Curiel and Dagonel (2020) for BISG use in turnout, Enos (2016) for BISG use in the study of local racial segregation, and Craig and Richeson (2018) for a study that considers perceptions of a person's local racial diversity on views of discrimination. A similar table for Florida appears as Table A.4 in the Supplementary Material.

FIGURE 3. 2016 Turnout Rates in North Carolina by Race and Census Tract Income for Self-Reported Race, BISG Model, and BISG + Random Forest Algorithm
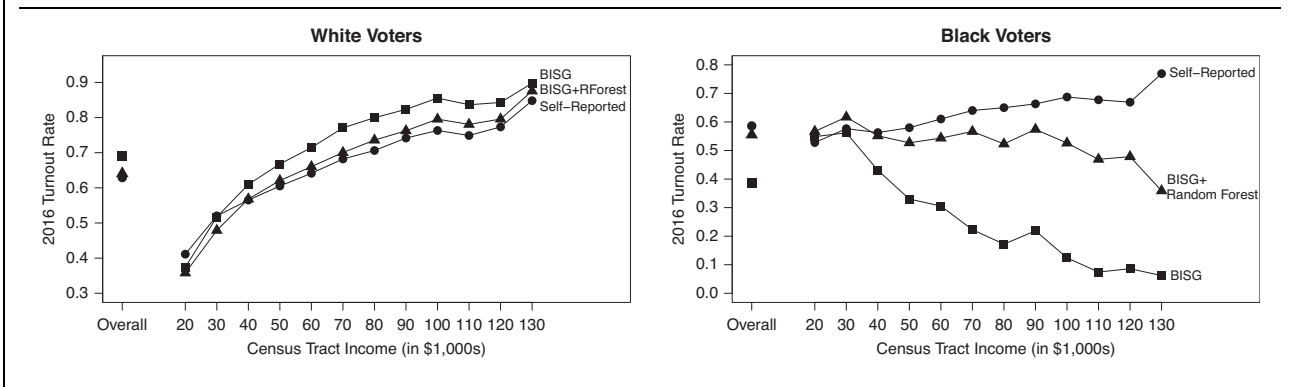
the citizen voting age population broken down by race to gauge the number of eligible voters by racial group (Fraga 2018).

Figure 3 illustrates how voter turnout estimates vary based on income and race for the different models. Among white voters (left panel), the BISG model overpredicts turnout by approximately 6 percentage points overall, whereas the BISG + random forest model is much closer (1.1 percentage points off overall) to the true turnout rate calculated using self-reported race. When looking across income levels, the differences between white voter turnout rates using the BISG estimates versus self-reported race grow in some cases to more than 10 percentage points. Among Black voters, the problem is even more concerning, both overall and when considered across income levels. Because the BISG model misclassifies so many Black voters in high-income neighborhoods, the associated turnout rate approaches zero, leading to an incredibly large underprediction of Black turnout overall. While a researcher using this method would hopefully recognize these implausible results at high levels of census tract income, they may not be so lucky if only considering Black turnout overall, and would, therefore, underestimate Black turnout by nearly 20 percentage points. This is a direct consequence of the BISG model's inability to make accurate predictions for high SES Black individuals. The BISG + random forest model, while missing the mark by a wide margin in the wealthiest census tracts (but not nearly to the degree as the BISG model alone), is much more accurate (3.1 percentage points off overall).

CONCLUSION

Documenting differences across races among a variety of political, economic, health, or policy outcomes is essential for not only understanding where inequalities exist, but also in crafting solutions that appropriately address disparities and improve outcomes for all people. However, such demographic information is not always available where we would like it, and many scholars use imputed predictions of race and ethnicity

where self-reported data are not available. Research relying on these methods needs to carefully account for the possibility of misclassification bias leading to incorrect estimates of explanatory relationships. To address this problem, we propose an ensemble method, where BISG predicted probabilities are incorporated into a random forest model along with other covariates. This approach substantially reduces the correlation between misclassification and various political and economic factors, especially for Black voters. The results presented here are important for scholars of many disciplines where political, economic, and health inequalities may exist across ethnicities and genders. We encourage researchers and practitioners to take seriously the likelihood that misclassification bias is correlated with core outcomes of interest, and to seek solutions—such as the ensemble model we propose—to mitigate such biases when using imputed race in policy studies.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055423000229.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available in the American Political Science Review Dataverse at https://doi.org/10.7910/DVN/FEOKT6.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human subjects.

REFERENCES

Adjaye-Gbewonyo, Dzifa, Robert A. Bednarczyk, Robert L. Davis, and Saad B. Omer. 2014. "Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study." *Health Services Research* 49 (1): 268–83.

Ansolabehere, Stephen, Bernard L. Fraga, and Brian F. Schaffner. 2020. "The CPS Voting and Registration Supplement Overstates Minority Turnout." *Journal of Politics* 84 (3): 1850–5.

Argyle, Lisa P., and Michael Barber. 2023. "Replication Data for: Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/FEOKT6

Baines, Arthur P., and Marsha J. Courchane. 2014. "Fair Lending: Implications for the Indirect Auto Finance Market." Study Prepared for the American Financial Services Association.

Craig, Maureen A., and Jennifer A. Richeson. 2018. "Majority No More? The Influence of Neighborhood Racial Diversity and Salient National Population Changes on Whites' Perceptions of Racial Discrimination." *RSF: The Russell Sage Foundation Journal of the Social Sciences* 4 (5): 141–57.

Curiel, John, and Angelo Dagonel. 2020. "Wisconsin Election Analysis." *Stanford-MIT Healthy Elections Project* 6: 2020–08.

Edwards, Frank, Michael H. Esposito, and Hedwig Lee. 2018. "Risk of Police-Involved Death by Race/Ethnicity and Place, United States, 2012–2018." *American Journal of Public Health* 108 (9): 1241–8.

Edwards, Frank, Hedwig Lee, and Michael 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): 16793–98.

Einstein, Katherine Levine, David M. Glick, and Maxwell Palmer. 2019. *Neighborhood Defenders: Participatory Politics and America's Housing Crisis*. Cambridge: Cambridge University Press.

Elliott, Marc N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole 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 (2): 69–83.

Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60 (1): 123–42.

Enos, Ryan D., Aaron R. Kaufman, and Melissa L. Sands. 2019. "Can Violent Protest Change Local Policy Support? Evidence from the Aftermath of the 1992 Los Angeles Riot." *American Political Science Review* 113 (4): 1012–28.

Fiscella, Kevin, and Allen M. Fremont. 2006. "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity." *Health Services Research* 41 (4p1): 1482–500.

Fraga, Bernard, and John Holbein. 2020. "Measuring Youth and College Student Voter Turnout." *Electoral Studies* 65: 102086. https://doi.org/10.1016/j.electstud.2019.102086

Fraga, Bernard L. 2018. *The Turnout Gap: Race, Ethnicity, and Political Inequality in a Diversifying America*. Cambridge: Cambridge University Press.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake News on Twitter during the 2016 US Presidential Election." *Science* 363 (6425): 374–78.

Grumbach, Jacob M., and Alexander Sahn. 2020. "Race and Representation in Campaign Finance." *American Political Science Review* 114 (1): 206–21.

Henninger, Phoebe, Marc Meredith, and Michael Morse. 2018. "Who Votes without Identification? Using Affidavits from Michigan to Learn about the Potential Impact of Strict Photo Voter Identification Laws." Working Paper.

Herring, Cedric, and Loren Henderson. 2016. "Wealth Inequality in Black and White: Cultural and Structural Sources of the Racial Wealth Gap." *Race and Social Problems* 8 (1): 4–17.

Hofstra, Bas, and Niek C. de Schipper. 2018. "Predicting Ethnicity with First Names in Online Social Media Networks." *Big Data & Society* 5 (1). https://doi.org/10.1177/2053951718761141

Imai, Kosuke, and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24 (2): 263–72.

Martino, Steven C., Robin M. Weinick, David E. Kanouse, Julie A. Brown, Amelia M. Haviland, Elizabeth Goldstein, John L. Adams, et al. 2013. "Reporting CAHPS and HEDIS Data by Race/Ethnicity for Medicare Beneficiaries." *Health Services Research* 48 (2pt1): 417–34.

Nguyen, Vy T., Ross D. Zafonte, Jarvis T. Chen, Kalé Z. Kponee-Shovein, Sabrina Paganoni, Alvaro Pascual-Leone, Frank E. Speizer, et al. 2019. "Mortality among Professional American-Style Football Players and Professional American Baseball Players." *JAMA Network Open* 2 (5): e194223. https://doi.org/10.1001/jamanetworkopen.2019.4223

Thomas, Timothy Andrew. 2017. "Forced Out: Race, Market, and Neighborhood Dynamics of Evictions." PhD diss. Department of Sociology, University of Washington.

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

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who Votes?* New Haven, CT: Yale University Press.