Supplemental Materials: Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records

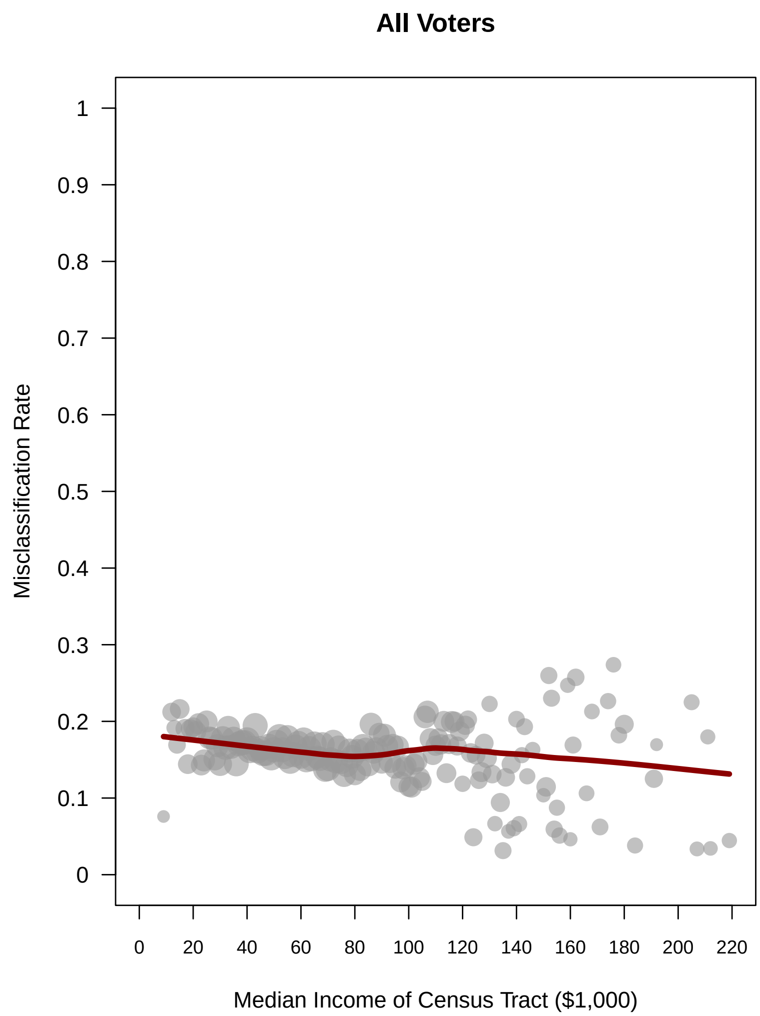
Lisa P. Argyle
Brigham Young University
lpargyle@byu.edu

Michael Barber
Brigham Young University
mbarber@byu.edu

December 27, 2022

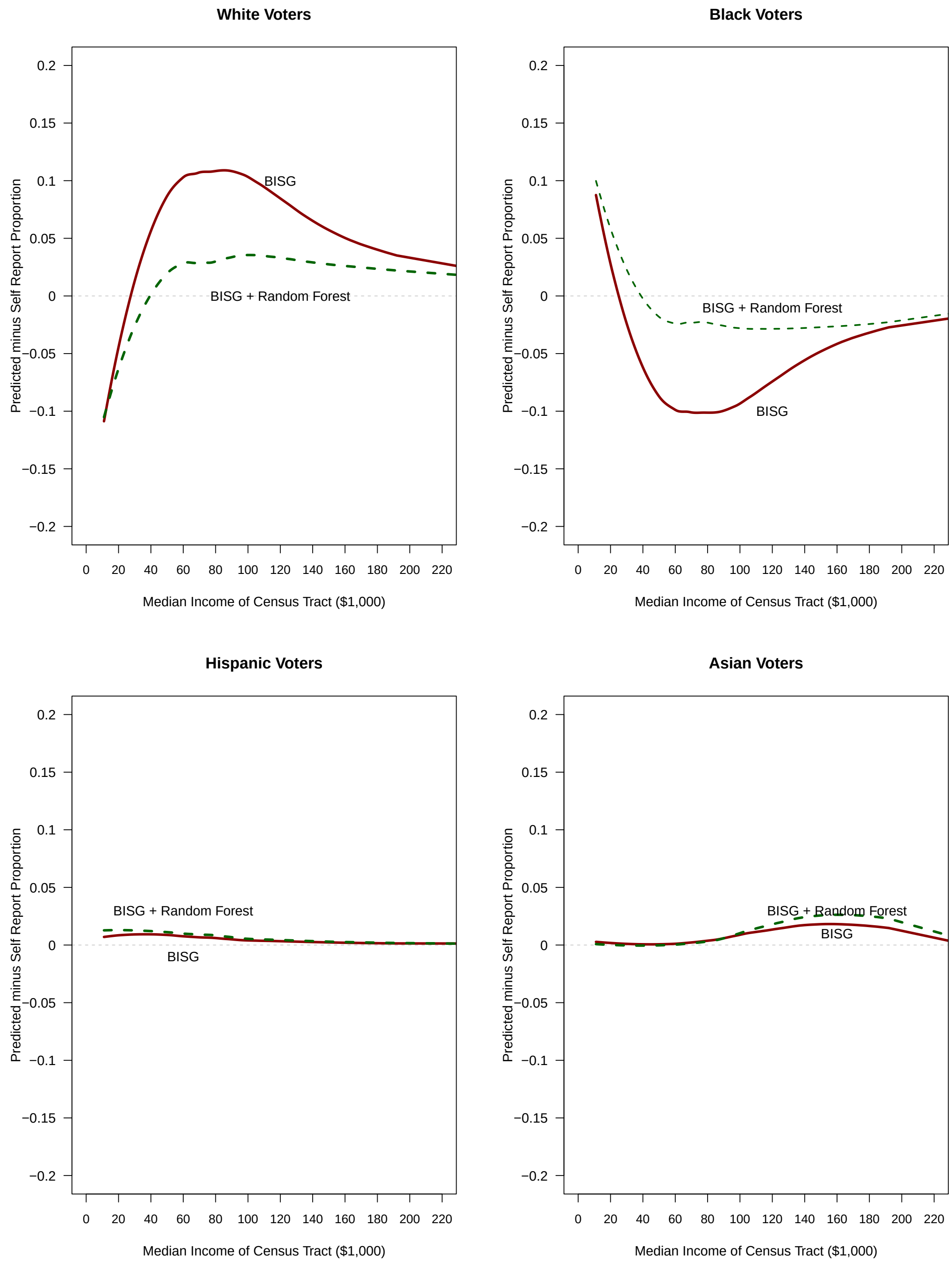
A.1 BISG Misclassification - Different Groups and Methods of Measuring Misclassification Rates

Figure A.1: **Misclassification Rates and Census Tract Income**



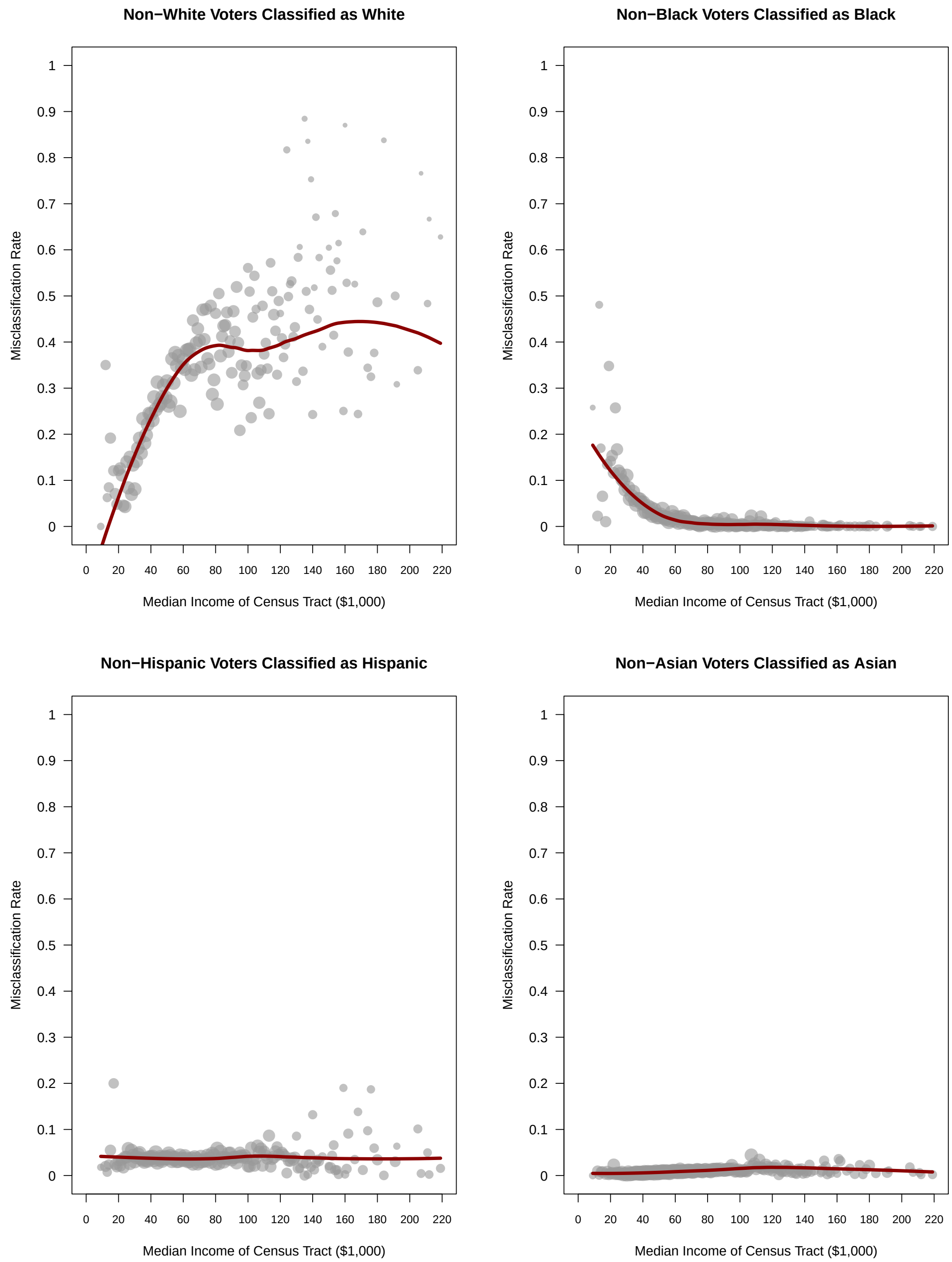
Note: This figure shows the relationship between voters' census tract income (x-axis) and the proportion of all voters, regardless of race, that are misclassified by the `wru` model. When we consider all voters together there is less of a relationship between census tract income and model misclassification rates. This is because the misclassification rate at lower incomes is higher for whites but lower for Blacks, which averages together to a moderate misclassification rate. Likewise, at higher tract income levels the high misclassification rate among Blacks and the low misclassification rate among whites averages together to a moderate misclassification rate overall. This illustrates the importance of considering the BISG model's performance for each racial group separately, which is displayed in Figure 1 of the main paper.

Figure A.2: **Net Error in Predicted Tract Racial Composition**



Note: Each panel shows the difference in the predicted racial composition of census tracts minus the actual composition of the census tract for each race using the BISG model and BISG + random forest model across tract income using a lowess line (span = 0.6) fit to the data. Values above zero indicate where the model has over-predicted tract composition of that racial group. Values below zero indicate where the model has under-predicted tract composition of that racial group. Values closer to zero indicate more accurate aggregate predictions by the model.

Figure A.3: **False Positive Rates by Census Tract Income**



Note: Each panel shows the relationship between voters' census tract income (x-axis) and the proportion of voters not of that race that are misclassified by the wru model as belonging to each race, using surname and census tract. Points, sized in proportion to number of observations, show average misclassification for each $1,000 increment. The line plots a lowess fit (span = 0.6) through those points, weighted by number of cases. The corresponding figure, showing false negative rates, is Figure 1 in the main paper.

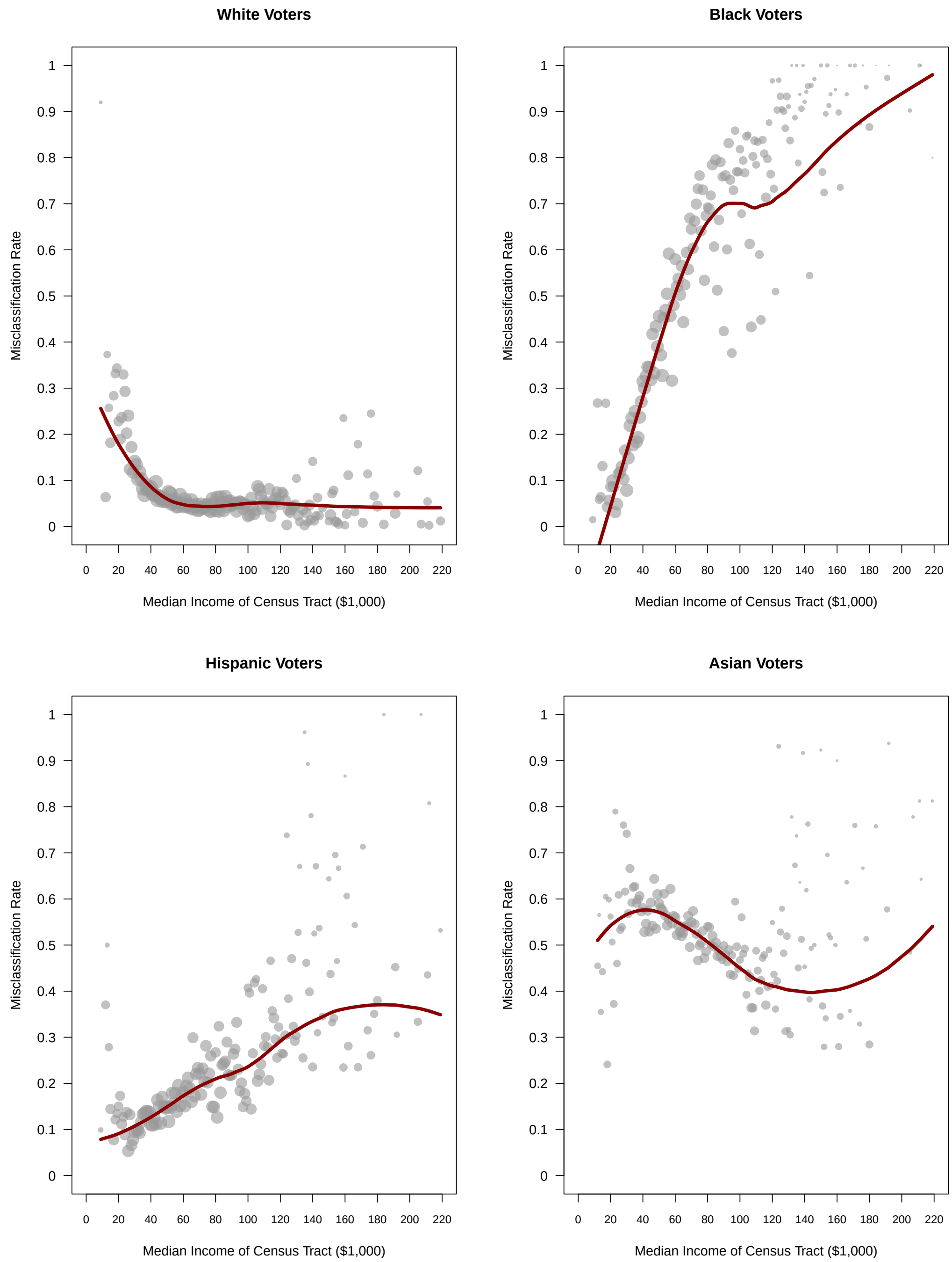
The Imai and Khanna (2016) BISG model allows for the addition of other variables aside from surname and census geography. Here we include additional variables in the BISG model: gender, and age. Table A.1 replicates Table 1 in the main paper and reports the overall error rates as well as false positive and false negative rates for models that add these other variables. The inclusion of this information does not substantially inform the model about a person's race.

Table A.1: **Replication of Imai and Khanna (2016) using 2018 Florida Voter File**

		Surname Census Tract	*Surname Census Tract Party*	*Surname Census Tract Party Gender*	*Surname Census Tract Party Gender Age*
Overall error rate:		0.163	0.151	0.149	0.151
White (63.2%)	False positive	0.287	0.247	0.245	0.261
	False negative	0.065	0.067	0.067	0.061
Black (13.4%)	False positive	0.025	0.029	0.029	0.025
	False negative	0.435	0.335	0.335	0.356
Hispanic (16.4%)	False positive	0.037	0.037	0.036	0.035
	False negative	0.146	0.151	0.152	0.164
Asian (1.9%)	False positive	0.009	0.008	0.008	0.008
	False negative	0.475	0.475	0.478	0.529
Other (2.6%)	False positive	0.000	0.000	0.001	0.001
	False negative	0.997	0.997	0.995	0.992

Note: Classification error rates by race and type of error. The false positive rate indicates the proportion of predicted cases in a race category that are incorrect (i.e. proportion of predicted whites who are not white). The false negative rate indicates the proportion of people who self-identify as a particular race who are misclassified (i.e. proportion of self-identified Hispanics who were classified as non-Hispanic). Numbers in parentheses in the first column represent the proportion of voters in the voter file who self-identify with each racial category. The second column of results is contained in Table 1 of the main paper. The first, third, and fourth columns show results for other combinations of demographic variables included in the BISG model: gender, and age. Rates of misclassification across racial groups are similar to those in Table 1 of the main paper.

Figure A.4: **Misclassification and Tract Income with BISG Model that includes party, gender, and age**



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, census tract, partisan registration, gender, and age. These figures correspond to Figure 1 in the main paper where the model includes only surname and tract. As can be seen, including the additional demographics of party, gender, and age from the voter file does not substantially change the patterns of bias observed in Figure 1. Points show average misclassification rate for each $1,000 increment. The line plots a weighted (by number of observations in each bin) lowess fit (span = 0.6) through those points.

A.2 BISG Misclassification Rates and Other Economic and Political Factors

Figure A.5 below looks at the relationship between various economic and political variables and the misclassification rate of the BISG model across the four main racial categories. Each column presents a different measure and each row presents results for a different racial group. The variables included are: proportion of a voter's census tract with a college degree, the proportion of a voter's census tract that are homeowners, a voter's propensity to vote in the 2016 presidential election, and the per capita campaign contributions made in each individual's zip code.[1] The rows of the figure show the misclassification rates for whites, Blacks, Hispanics, and Asians, respectively. Histograms above each figure show the distribution of the data across each variable for each racial group.

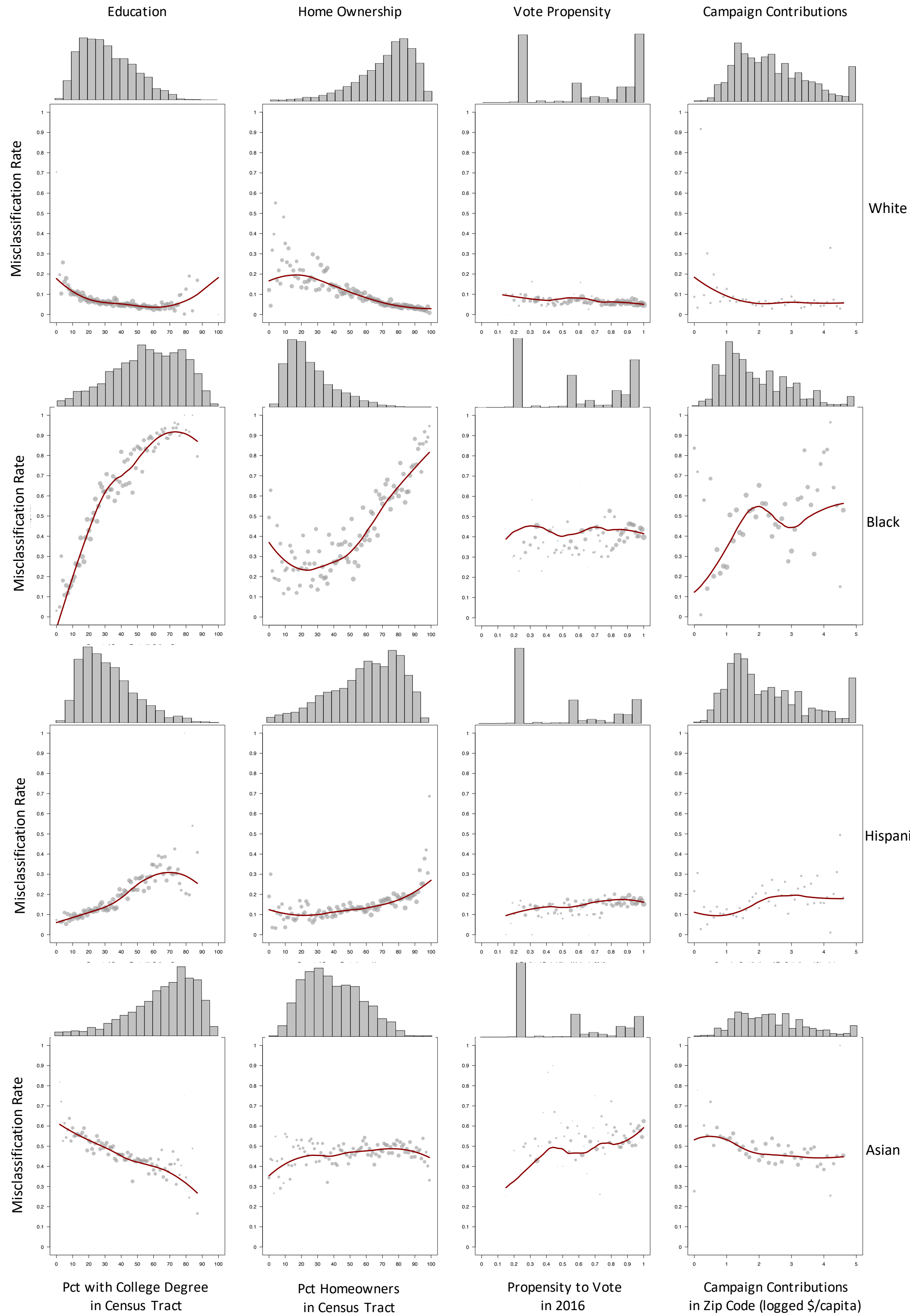
The relationship between these variables and the misclassification rate are largely similar to those shown in Figure 1 in the main paper. Collectively, they show that Blacks who reside in the most socio-economically well-off neighborhoods (highest median income, most educated, highest homeownership, and most campaign contributions per capita) are incredibly likely to be misclassified as non-Black by the racial prediction model. Interestingly, there is not a significant relationship between the propensity to vote and model misclassification among Blacks, but there is one across the other three racial groups, albeit smaller than the other factors considered in Figure A.5.

The BISG model also struggles most when a person lives in a neighborhood in which they are a racial minority. This is true among all races in Florida and North Carolina, especially among white and Black individuals. To show this we generate the racial composition of each voter's census tract and then plot the misclassification rate for each voter and compare that race across the racial diversity of the census tracts. Figure A.6 shows these results in both Florida (top row) and North Carolina (bottom row). Across all races, people who live in census tracts with few other co-ethnic individuals are dramatically more likely to be misclassified by the BISG model.

When a person lives in a census tract in which their race comprises fewer than 20% of the population, the proportion of people of that race who are misclassified by the BISG model is often greater than 65%. This is especially true among white and Black voters. The model performs better overall among Hispanic individuals, even when those individuals live in census tracts with very few Hispanics. In North Carolina there are very few tracts with large proportions of Hispanics or Asians. In Florida, this is also true for census tracts with large proportions of Asians. In both states there are census tracts that span essentially all possible proportions of whites and Blacks (i.e. from 0 to 100% of each race). The fact that the BISG model performs especially poorly for individuals who are minorities in their local communities presents difficult challenges for scholars who would use the BISG model to study racial segregation, polarization, and the causes and consequences of spatial sorting on race. Table 2 in the main paper shows how this systematic error leads to the under-prediction of racial segregation. Using the modeled race variable, a smaller proportion of individuals of all races are estimated to be "local minorities", or people living in census tracts where their race is not the most common ethnic group. When looking instead at the self-reported race, we see a higher level of local minorities. Thus, for scholars of residential segregation, the BISG model will systematically under-represent the number of people who are minorities in their communities and overstate residential segregation.

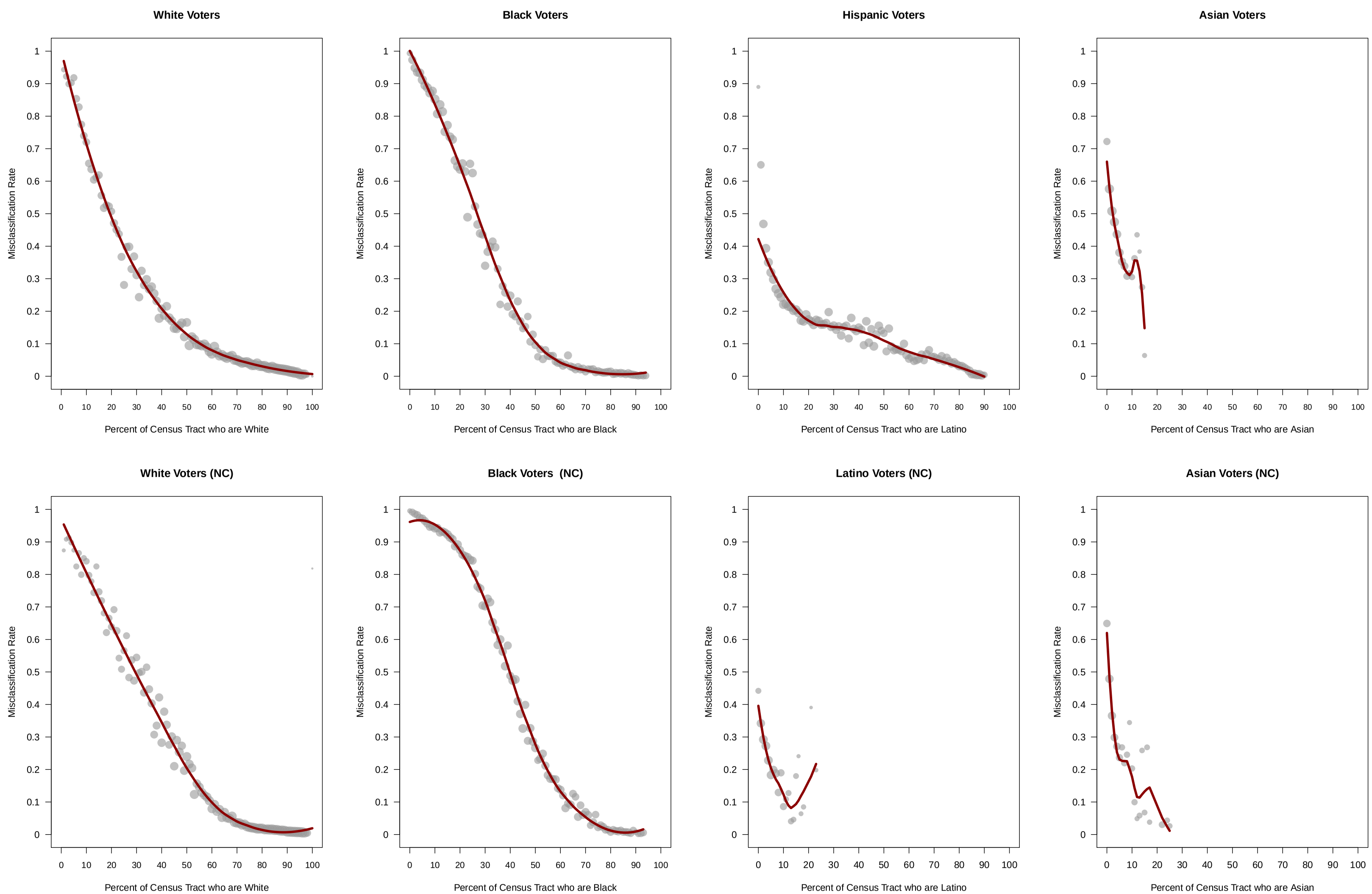
[1]Campaign contributions are aggregated at the zip code level rather than the census tract level because of how contributions are reported to the FEC.

Figure A.5: **Misclassification Rates by Race and Economic and Political Variables, Florida**



Note: Each panel shows the relationship between demographics (x-axis) and the proportion of voters from each race that are misclassified by the wru model (y-axis). Panel columns show levels of education, home ownership, vote propensity, and campaign contributions per capita. Rows show misclassification rates for white, Black, Hispanic, and Asian individuals.

Figure A.6: **Misclassification Rate and Census Tract Racial Composition in Florida (top row) and North Carolina (bottom row)**



Note: Each panel shows the relationship between voters' census tract racial composition (x-axis) and the proportion of voters from that race who are misclassified by the wru model. For example, the top left panel shows the proportion of white voters who are misclassified based on the proportion of white voters in that person's census tract. The pattern in all cases is that voters who live in census tracts where they are a racial minority (the left side of the x-axis) are much more likely to be misclassified by the wru model. Points show average misclassification rate for each .01% bin. The line plots a weighted (by number of observations in each bin) lowess fit (span = 0.6) through those points.

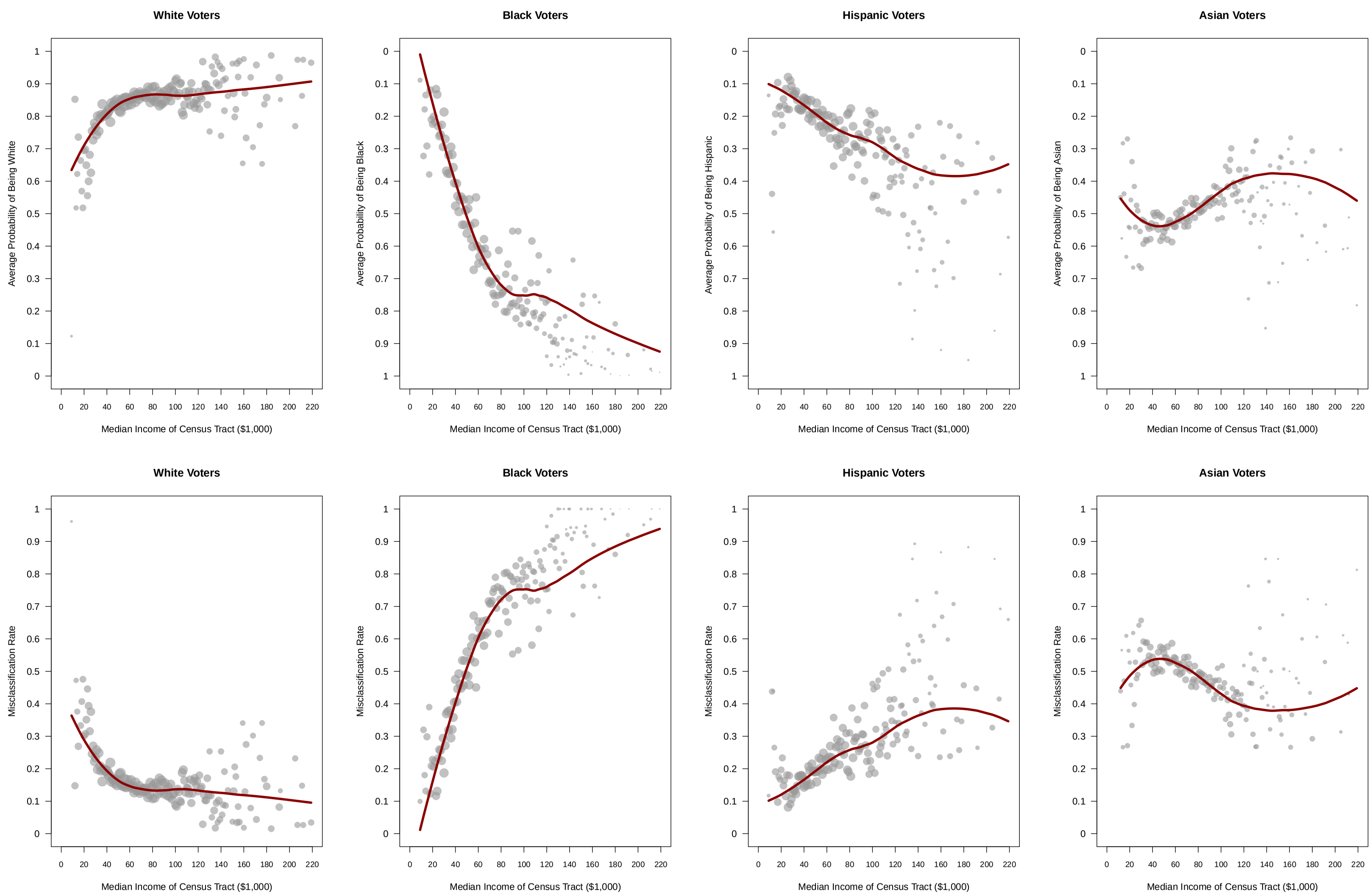
A.2.1 Alternative Methods of Race Classification

In the main paper we classify a voter's predicted race by choosing the race that is assigned the highest probability by the BISG model. Here we consider two alternative methods of assigning predicted race to see if the correlations between race misclassification and various SES factors like census tract income are lower using these alternative methods.

The first method uses the posterior probability distribution generated by the BISG model across each racial category and does not make a binary prediction of race but instead calculates the average predicted probability of belonging to each racial group among all self-reported people of that particular race. For example, the top left panel of Figure A.7 shows the average probability assigned to being white for all self-reported white voters across different census tract income values. The pattern is largely the same as what is shown in the main paper, just inverted (lines slope down instead of up among Black voters) since the y-axis is now measuring the probability of accurate prediction rather than misclassification.

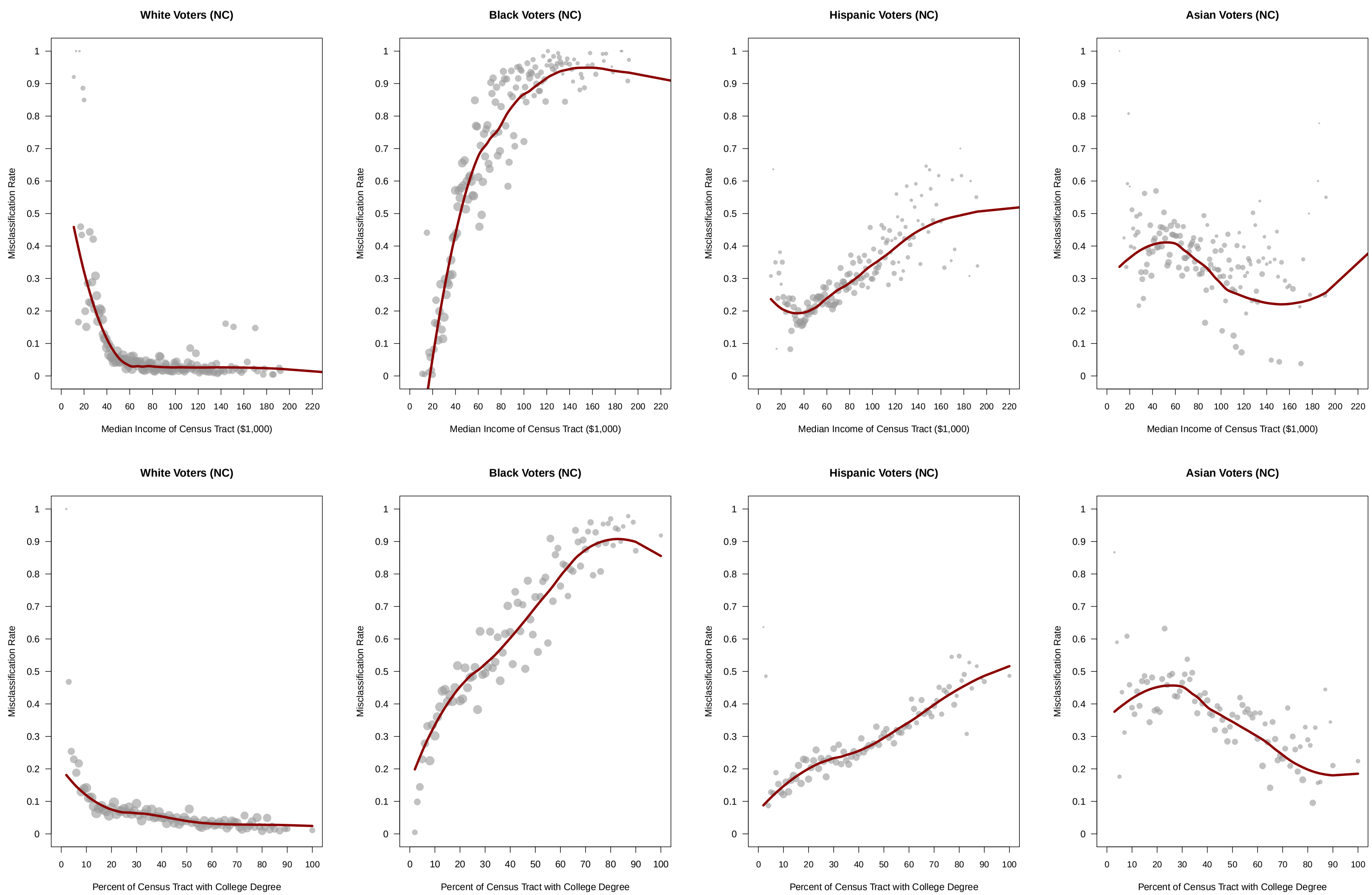
The second method generates a binary predicted race based on the posterior probability distribution across races generated by the BISG model. However, rather than classifying a voter's predicted race by taking the category with the highest probability, we select a predicted race as a draw from a five-category discrete probability distribution (white, Black, Hispanic, Asian, Other) using the distribution of probabilities across all racial categories. Thus, the predicted race is stochastic rather than deterministic. The results of this classification method and the relationship between misclassification and census tract income are shown below in the bottom row of Figure A.7 and are largely the same as those reported in the main paper in Figure 1.

Figure A.7: **Alternative Methods of Predicting Race: Average Probability (top row) and Draw from Probability Distribution (bottom row) - Florida**



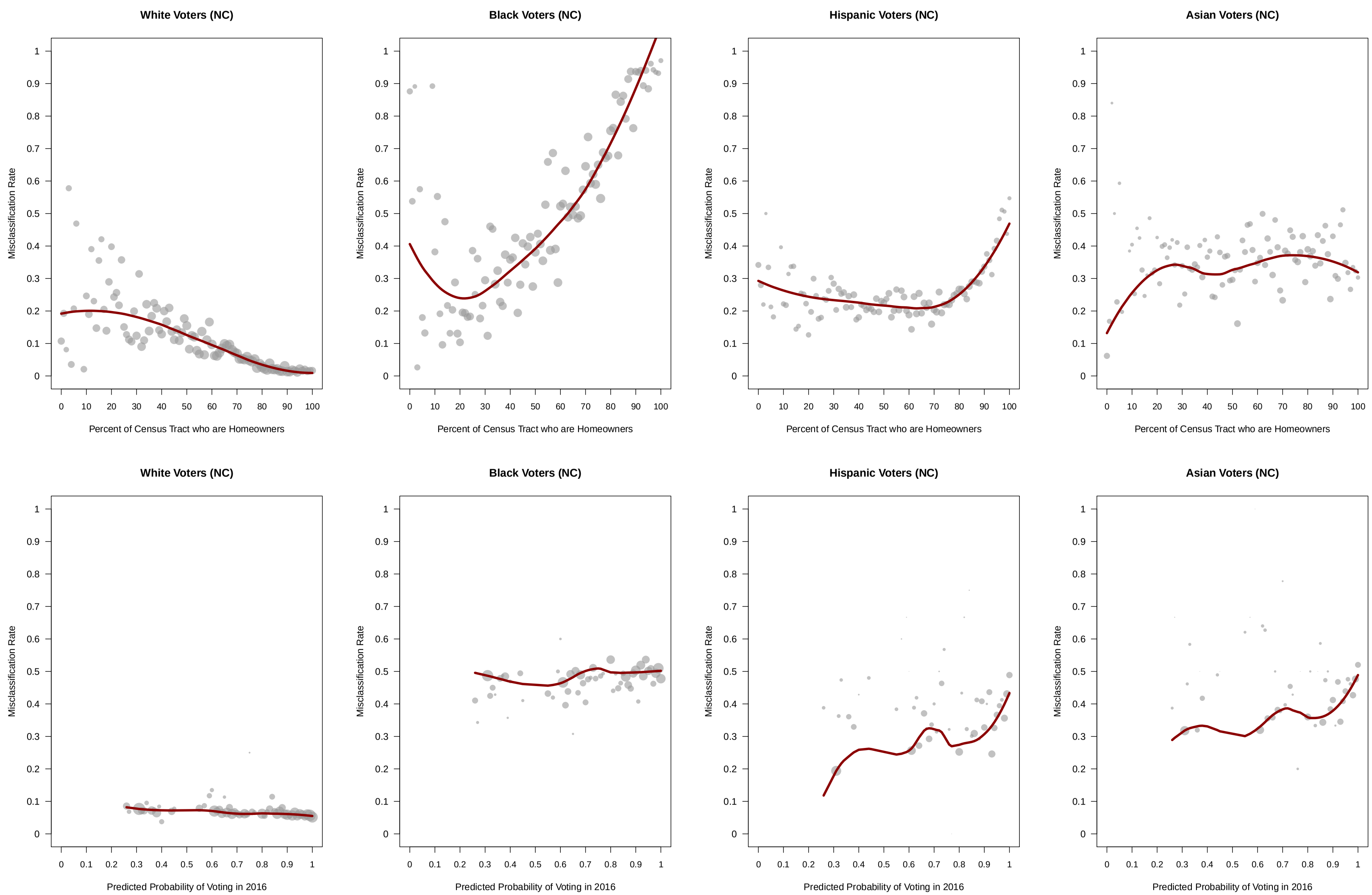
Note: Each panel in the top row shows the relationship between voters' census tract income (x-axis) and the average probability that voters of a particular race are predicted by the `wru` model to be of that race. In the wealthiest census tracts the probability of being Black (among self-reported Black voters) is nearly zero. In the bottom row predicted race is calculated by taking a draw from a discrete probability distribution with probabilities equal to the posterior probabilities generated by the `wru` model instead of simply assigning the predicted race to be the category with the highest probability (as is the case in Figure 1 in the main paper). The results, however, are nearly identical regardless of method.

Figure A.8: **Misclassification Rate and Census Tract Income (top row) and Education (bottom row) - North Carolina**



Note: Data from North Carolina voter registration file. Each panel in the top row 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. The bottom row shows the relationship between voters' census tract percent with a college degree (x-axis) and the proportion of voters from each race that are misclassified by the wru model. Points show average misclassification rate for each $1,000 increment. The line plots a weighted (by number of observations in each bin) lowess fit (span = 0.6) through those points.

Figure A.9: **Misclassification Rate and Census Tract Home Ownership (top row) and Vote Propensity (bottom row) - North Carolina**



Note: Data from North Carolina voter registration file. Each panel in the top row 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. The bottom row shows the relationship between voters' census tract percent with a college degree (x-axis) and the proportion of voters from each race that are misclassified by the wru model. Points show average misclassification rate for each $1,000 increment. The line plots a weighted (by number of observations in each bin) lowess fit (span = 0.6) through those points.

A.3 Gender

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 primarily 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 particular minority women are especially difficult to correctly classify. Figure A.10 shows that this is the case. The left panel shows the misclassification rate of the model by gender and race. We see that Latina and Asian women are 6 and 15 percentage points more likely to be misclassified than their male counterparts, respectively. On the other hand, there are much smaller differences in misclassification among white and Black men and women. The right panel of Figure A.10 shows rates of interracial marriage in 2017. We see that Latina and Asian women — the two groups where misclassification rates are most different from their male counterparts — are also the two groups most likely to be married to a person of a different race. And while other factors certainly also contribute to the misclassification rates, the differences observed here are suggestive of gendered differences in the ability of the model to accurately estimate a person's ethnicity.

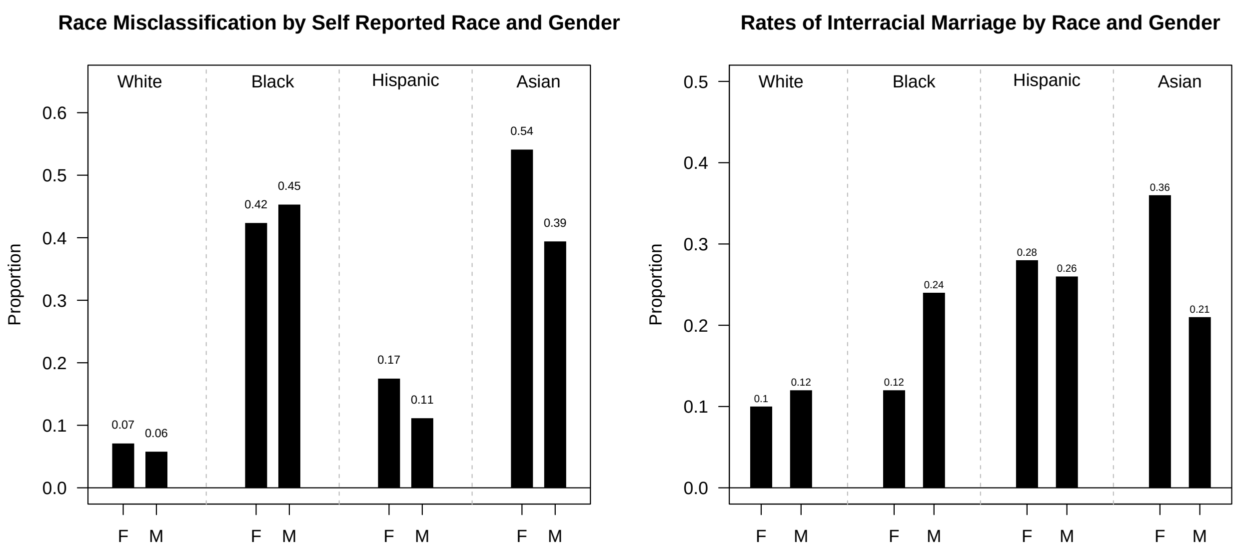


Figure A.10: The left panel shows misclassification rates by race and gender. The right panel shows rates of interracial marriage by race and gender. Marriage data is from the Pew Research Center: https://www.pewsocialtrends.org/2017/05/18/1-trends-and-patterns-in-intermarriage/

A.4 Technical Discussion of Random Forest Estimation

A.4.1 Overview

In order to reduce the correlation of misclassification with other demographic variables, we propose an ensemble model that incorporates the BISG predicted probabilities of each racial classification into a random forest model that accounts for a variety of individual- and neighborhood-specific factors and adjusts for class imbalance. A random forest is an aggregated collection of classification trees,[2] where at each decision split in the tree only a random subset of variables are available for selection. Random forests are widely used in both data science and social science applications and are appropriate for use with unordered multi-class outcomes. In this appendix, we describe the estimation procedures for the random forest model presented in the paper. Upon publication acceptance, we will publicly post the random forest model object with the replication materials and on the author's GitHub, so that future researchers with comparable data can use it for racial classification predictions without needing to re-train their own model. We provide an example of this with an out-of-sample prediction in North Carolina using the model that was trained with data in Florida.

A.4.2 Data

Starting with the complete 2018 Florida Voter File, we partitioned the data into three separate sets for training (60%; n = 8,448,979), validation (20%; n = 2,816,327), and testing (20%; n = 2,816,326). We used the 60% sample as training data for the random forest models. We used the 20% validation sample in the hyper-parameter tuning process. The final 20% of the data is the test set for all accuracy comparisons reported in the paper, meaning the same 20% set of cases is used in all calculations that compare the performance of BISG to the random forest in the Florida data. We only compare out-of-sample predictions so as to avoid overfitting the data, which would mean producing overly accurate or confident predictions based on the unique aspects of the training data that do not generalize to other samples. In all three data partitions, cases in the original data file with missing data in any of the predictors or an "Unknown" classification for the self-reported race outcome are dropped from the analysis and are not predicted. This results in training (n = 8,051,010), validation (n = 2,683,479), and testing (n = 2,684,084) sets. The outcome variable is self-reported race, with five unordered category options: White, Black, Hispanic, Asian, or Other. The complete set of predictors for the random forest is:

- BISG Probability White
- BISG Probability Black
- BISG Probability Hispanic
- BISG Probability Asian
- BISG Probability Other Race
- Party
- Sex
- Age
- Median Census Tract Income, rounded to nearest $1,000
- Percent of Census Tract with College Education, rounded to nearest 1%

[2] A classification tree is an algorithmic process by which an outcome space is "split" into progressively smaller "nodes," or sets of cases, by selecting the binary division of available predictors that most improves the correct classification of cases in the training data.

- Predicted Probability of Individual Voting
- Population in Zip Code
- Donations Per Capita, rounded to nearest 1%
- Respondent is a Campaign Donor, binary indicator
- CVAP (Citizen Voting Age Population) % in Census Tract who are White
- CVAP % in Tract who are Black
- CVAP % in Tract who are Hispanic
- CVAP % in Tract who are Asian
- CVAP % in Tract who are Other Race
- Percent of Tract who are Homeowners, rounded to nearest 1%
- Median Census Tract House Price, rounded to nearest $10,000

The first five variables are the probabilities generated by the BISG model. In other words, we first run the Imai and Khanna (2016) model to generate the predicted probabilities associated with each racial group. We then include those probabilities as variables in the random forest models. Because random forests can be biased towards variables that have more potential splits, we use the rounded version of several census-tract level variables rather than the exact figures. All of these variables are derived from publicly available data: the BISG probabilities can be derived from surname and location data available in a voter file or other data source; political party, sex, and age are available from public or commercial voter files; neighborhood socioeconomic, racial, and population data are available from the Census Bureau; and campaign donation histories are available from the FEC.

A.4.3 Class Imbalance

Class imbalance can pose a significant challenge for classification tasks. Because one class (in this case, "White") is much more common than other classes in the training data, probabilistically the majority class is the best choice when there is uncertainty over class prediction. This is particularly the case when percent accuracy is the optimized metric. This can result in the major class being predicted at an artificially high rate. In the BISG race predictions, this is apparent in the relatively high false positive rate for white classification. Because there is significant class imbalance in the data, and because the over-classification of racial and ethnic minority voters as white is a central contributing factor in the bias we document in this paper, some adjustment for class imbalance is essential. We take two approaches to compensating for class imbalance in the random forest estimation: class weights and optimization using the F1-Score.

First, we provided class weights in the random forest algorithm. The class weights adjust the calculation of the best split at each tree node to place extra emphasis on correct classification of cases in minority classes. The class weights we used are the square root of the inverse proportion of each racial group in Florida, based on 2018 Census Bureau estimates. See Table A.2 below for the population proportions and final weights.

Table A.2: **Florida Population Distribution and Class Weights**

Racial/Ethnic Group	Population Percent	Class Weight
White (non-Hispanic)	53.26	1.37
Black	15.27	2.56
Hispanic	26.12	1.96
Asian	2.77	6.01
Other	2.55	6.26

Moreover, because biases from class imbalance can be exacerbated by the use of percent accuracy as an evaluation metric, we used the alternate metric of an F1-Score to select the optimal model. The F1-Score is a measure of model accuracy based on a combination of precision and recall. Precision is the fraction of cases correctly classified to a class out of all cases assigned to that class; recall is the fraction of cases correctly classified to a class out of all cases that are actually in that class. We use an unweighted macro-averaged F1-Score, which means that the F1-Score is first calculated separately for each class and then averaged across the five classes to produce a single global metric of model performance. This approach means that both correct and incorrect predictions are used as part of the evaluation metric, and that the performance in each racial group is weighted equally to performance in every other racial group, without regard to the size of the racial group.

A.4.4 Hyperparameter Tuning

Using a grid search approach,[3] we ran 24 iterations of the random forest model to find the best combination of the following hyper-parameters: split rule,[4] number of variables sampled as candidates at each node split ("mtry"),[5] and minimum final node size.[6] Each of the 24 random forest models was trained on the 60% training data set, and then validated using the 20% validation set. We use F1-Score as the evaluation metric to select the best-performing set of hyper-parameters. Therefore, the best model is the model with the highest F1-Score based on the out-of-sample predictions made on the 20% validation set (note that this is separate from the 20% test set reported in the main text). We also report the total percent accuracy and Cohen's Kappa metrics for each combination of parameters. Table A.3 below presents the hyper-parameter specifications and performance metrics for each of the 24 random forest estimations, sorted by F1-Score. The highest performing model for each metric (F1-Score, total percent accuracy, and Cohen's Kappa) is bolded.

[3]A grid search approach means that every combination of selected hyperparameters is estimated, and the best model is chosen from that exhaustive set.

[4]When selecting each split in the classification tree, different metrics can be used to identify the split that provides the most improvement in classification. Here we use two standard options, both of which are available in the ranger R package: "gini" and "extremely randomized trees."

[5]The "random" component of a random forest is introduced by randomly selecting only a subset of the total predictors that can be used at each split in each tree. This parameter can vary from 1 to the total number of predictors. In this case, we test values of 2, 3, and 4

[6]Minimum node size refers to when the random forest stops partitioning the data. This can be as low as 1, meaning the algorithm continues until there is only one observation is left in each subdivided classification space (or "node", sometimes also called "leaves.") "Trimming" the trees, by increasing the final node size, can be one way to prevent overfitting the model. We vary this parameter from 10 to 25 in increments of 5.

Table A.3: **Hyperparameter Tuning Performance Metrics**

Model Number	mtry	Split Rule	Node Size	F1	Accuracy	Kappa
17	4	gini	10	**0.628**	0.857	0.725
18	4	gini	15	0.627	0.857	0.725
9	3	gini	10	0.627	**0.857**	0.725
10	3	gini	15	0.627	0.857	**0.726**
19	4	gini	20	0.627	0.856	0.725
20	4	gini	25	0.626	0.856	0.725
11	3	gini	20	0.626	0.857	0.725
12	3	gini	25	0.626	0.856	0.725
1	2	gini	10	0.626	0.857	0.725
2	2	gini	15	0.625	0.857	0.725
3	2	gini	20	0.625	0.857	0.725
4	2	gini	25	0.624	0.856	0.725
21	4	extratrees	10	0.623	0.855	0.723
22	4	extratrees	15	0.623	0.855	0.723
13	3	extratrees	10	0.623	0.855	0.723
14	3	extratrees	15	0.622	0.855	0.723
23	4	extratrees	20	0.622	0.854	0.722
24	4	extratrees	25	0.622	0.854	0.722
15	3	extratrees	20	0.621	0.855	0.722
16	3	extratrees	25	0.621	0.854	0.722
5	2	extratrees	10	0.621	0.856	0.723
6	2	extratrees	15	0.620	0.855	0.722
7	2	extratrees	20	0.620	0.855	0.722
8	2	extratrees	25	0.619	0.855	0.722

A.4.5 Final Model Specifications

The specifications of the final model used for prediction and evaluation in the paper are as follows:

- Number of Trees: 250
- Number of Variables Sampled at Each Split: 4
- Minimum Node Size: 10
- Split Rule: Gini
- Class Weights Included
- Number of Predictors: 20
- Trained on: 60% Training data set

This final model was used to predict the racial classification of the 20% test data, as well as the full voter file from North Carolina. The classification rates for the 20% test data in Florida and the full North Carolina voter file are reported in Table 1 of the main paper. Figures 2 and A.12 show the misclassification rates from the random forest model across census tract income levels and compare the misclassification rates to the original BISG model predictions in North Carolina and Florida, respectively. As discussed in the main paper and below, there is dramatic improvement in classification, particularly among Black individuals in both states. Finally, Table A.4 shows the results of various summary statistics calculated across racial groups when predicting race using the BISG model and the random forest model and compares those estimates to the self-reported truth. As can be seen in the table, the random forest model produces estimates that are closer to the truth than the original BISG model in 67% of the calculations we estimate.

A.4.6 Computing Hardware and Statistical Software

The hyper-parameter tuning process was conducted using University supercomputing resources. We used 16 processors, each allocated with 256 GB of memory, and the estimation took approximately 47 hours to complete. All analysis was conducted in R using the following packages:

- Kuhn, M. (2008). “Building Predictive Models in R Using the caret Package.” Journal of Statistical Software, 28(5), 1 - 26. doi: `http://dx.doi.org/10.18637/jss.v028.i05`

- Kuhn, M., and D. Vaughan. (2021). “yardstick: Tidy Characterizations of Model Performance.” Version 0.0.8. `https://CRAN.R-project.org/package=yardstick`

- Silge, J., F. Chow, M. Kuhn, and H. Wickham. (2021). “rsample: General Resampling Infrastructure.” Version 0.1.0. `https://cran.r-project.org/package=rsample`

- Wright MN, Ziegler A (2017). “ranger: A Fast Implementation of random forests for High Dimensional Data in C++ and R.” Journal of Statistical Software, 77(1), 1–17. doi: `10.18637/jss.v077.i01`.

Table A.4: **Differences in Summary Statistics for Self-Reported Race versus Predicted Race using BISG and BISG + Random Forest models in Florida**

		Median Income	*Median Home Value*	*Campaign Donors*	*2016 Turnout Percent*	*Minority in Own Tract*
White:	Self-Reported	$62,926	$251,363	79.02%	65.78	7.11%
	BISG Model	$63,139	$250,640	81.35%	70.79	6.00%
	% Diff.	0.34	-0.29	2.95	7.62	-15.66
	BISG + RF Model	$63,276	$252,071	79.60%	67.00	5.64%
	% Diff.	0.56	0.28	0.73	1.86	-20.67
Black:	Self-Reported	$48,706	$206,590	6.14%	53.31	59.46%
	BISG Model	$42,287	$194,162	4.22%	38.08	36.27%
	% Diff.	-13.18	-6.02	-31.34	-28.57	-38.99
	BISG + RF Model	$47,423	$202,444	6.34%	54.88	56.14%
	% Diff.	-2.63	-2.01	3.30	2.93	-5.59
Hispanic:	Self-Reported	$58,247	$272,308	11.67%	45.21	51.60%
	BISG Model	$57,634	$270,257	12.83%	47.03	50.48%
	% Diff.	-1.05	-0.75	9.90	4.03	-2.15
	BISG + RF Model	$57,938	$272,030	12.41%	46.28	50.30%
	% Diff.	-0.53	-0.1	6.31	2.38	-2.52
Asian:	Self-Reported	$67,678	$260,957	1.47%	46.73	100.00%
	BISG Model	$69,474	$269,780	1.50%	44.37	100.00%
	% Diff.	2.65	3.38	2.57	-5.05	0.0
	BISG + RF Model	$70,233	$268,689	1.42%	40.62	100.00%
	% Diff.	3.78	2.96	-3.22	-13.08	0.0

Note: This table replicates the results of Table 2 (which uses data from NC) in the main paper using data from a randomly withheld 20% of the Florida voter file where we conduct an out of sample prediction based on the BISG + random forest model trained using the other 80% of the Florida voter file. Using predicted race based on the BISG model alone leads to estimates of lower median income and home value; lower rates of campaign donations; and lower rates of living as a minority in one's census tract among Black individuals. Incorporating our proposed solution of using the BISG + random forest correction improves estimates dramatically for this group. 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.

Figure A.11: **Reduction in Misclassification Error from Random Forest Model in Florida**

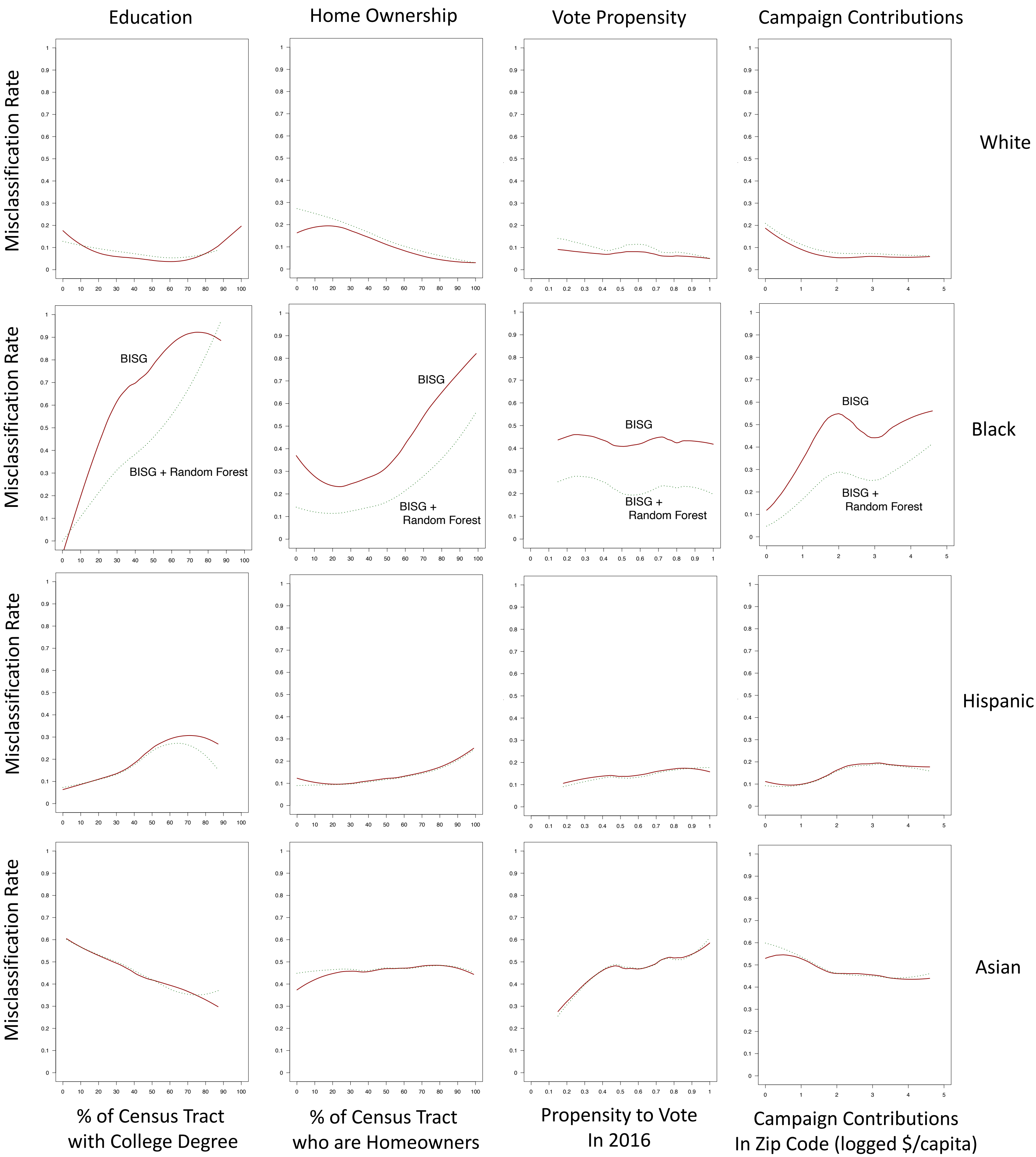
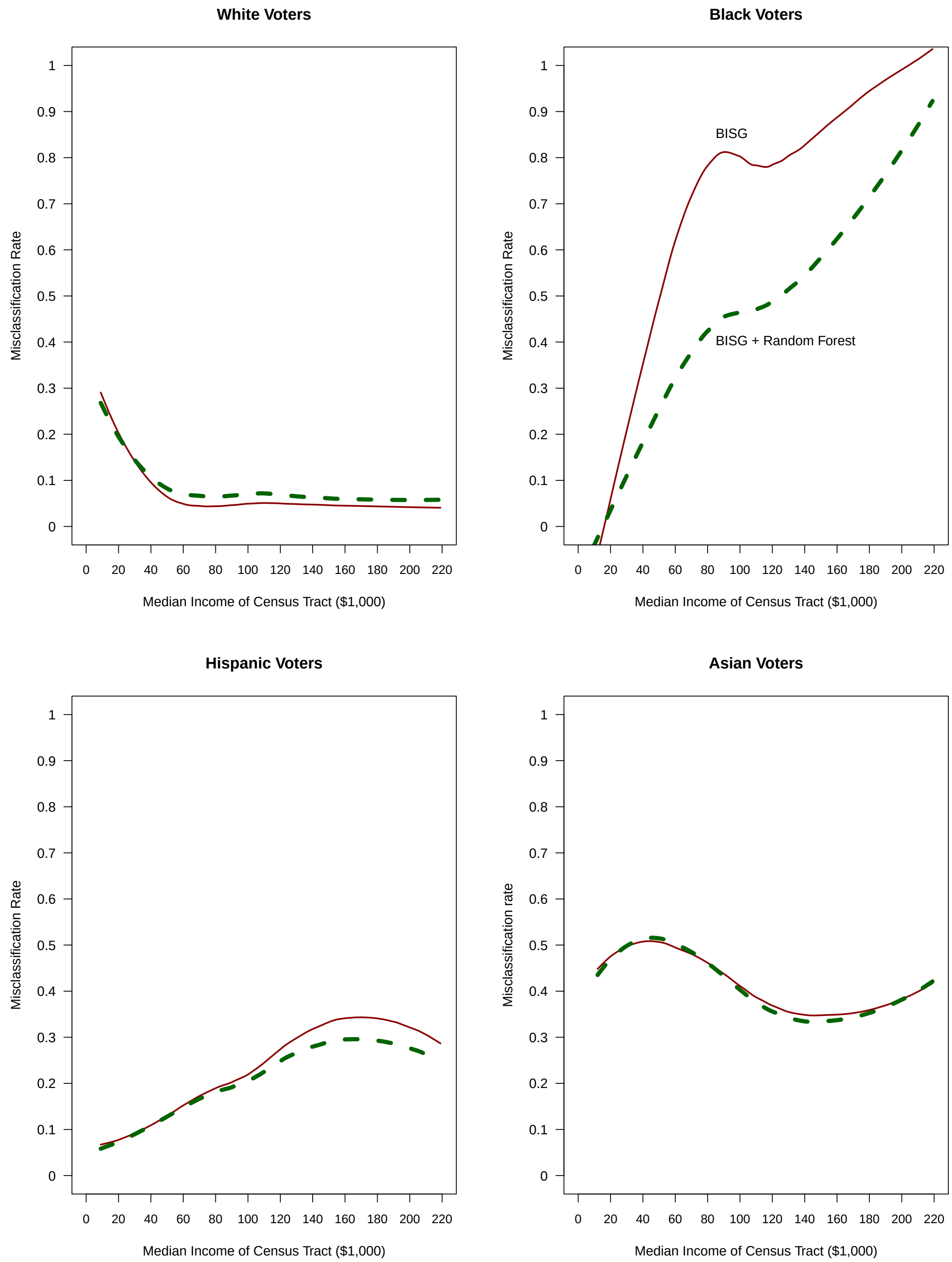


Figure A.12: **Misclassification Rates and Census Tract Income for BISG Model and BISG + random forest Algorithm in Florida**



Note: This figure replicates the results of Figure 2 of the main paper using data from Florida rather than North Carolina. The solid red line shows the average misclassification rate for the BISG model and the green dotted line shows the misclassification rate for the BISG model plus the random forest algorithm. Especially among Black voters, the addition of the random forest algorithm dramatically decreases misclassification rates.