Supplementary Material for "Validating the Applicability of Bayesian Inference with Surname and Geocoding to Congressional Redistricting"

A BISG Technical Appendix

Bayesian Improved Surname Geocoding (BISG) uses an individual's surname and location to estimate the probability that they are of a given racial category. As shown in Imai and Khanna (2016), BISG uses Bayes Theorem to calculate predicted probabilities for each individual's race:

$$P(R_i = r|S_i = s, G_i = g) = \frac{P(G_i = g|R_i = r)P(R_i = r|S_i = s)}{\sum_{i=1}^{n} P(G_i = g|R_i = r)P(R_i = r|S_i = s)}$$

where r indicates a racial group, s is individual i's surname, and g indicates some geographic location such as county, ZIP code, or Census block. Using the R package **zipWRUext** from Clark, Curiel, and Steelman (2021) on a voter file that includes voter surnames and ZIP codes outputs $P(R_i = r|S_i = s, G_i = g)$ for each voter – the predicted probability of individual i being of race r, given their surname s and geographic location g, where the set of G_i is estimated from the racial proportions within all ZIP codes present within the state. The posterior takes the form of the predicted probability $P_{i,r}$, where i indexes individuals and r indicates race (White, Black, Hispanic, Asian, or other). These categories are mutually exclusive.

Figure S1 illustrates the format of the surname dictionary present within WRU, specifically for the top 10 similar names for the white and Black racial categories. Each row sums to one, and all but one of the names are almost certainly white or Black. This is illustrative of a larger issue within American politics, where BISG performs worst in separating white and Black individuals given the legacy of slaveholder surnames. Situations like these are where the geographic demographic context becomes instrumental in preventing a noisy estimate for an individual's race.

surname	p_whi	p_bla	p_his	p_asi	p_oth
LUCKY	0.4360	0.4360	0.029100000	0.062100000	0.036900000
OMARI	0.3966	0.3966	0.067200000	0.020200000	0.119400000
DYCE	0.4742	0.4742	0.010300000	0.005200000	0.036100000
TENN	0.2595	0.2595	0.023850000	0.233300000	0.223850000
LANTON	0.4573	0.4573	0.038600000	0.006850000	0.039950000
WORLD	0.4613	0.4613	0.031000000	0.012350000	0.034050000
AUSBON	0.4757	0.4757	0.007300000	0.007300000	0.034000000
MACARTHY	0.4656	0.4656	0.012366667	0.031700000	0.024733333
BEMBRIDGE	0.4889	0.4889	0.011100000	0.011100000	0.000000000
FREEMEN	0.4464	0.4464	0.011900000	0.011900000	0.083400000
COES	0.4752	0.4752	0.016533333	0.016533333	0.016533333

Figure S1. Top 10 Similar White and Black Surnames from WRU 2010 Surname Dictionary

There are two intuitive ways one might use these BISG predicted race probabilities in redistricting. The first is the individual-level plurality assignment method, which we abbreviate as PM. This method assigns to each individual the race with the highest predicted probability. Once each voter has been assigned a race, an estimate of the racial composition of a precinct could be calculated by adding up the number of individuals of each race and dividing by the total number of individuals

in each precinct. Let $\alpha_{r,p}$ indicate the proportion of individuals in precinct p who are of race r. Formally, PM assigns a race r to individual i by solving for each voter:

$$r_i = \arg\max_r [P_{i,r}]$$

and $\alpha_{r,p}$ is calculated as:

$$\alpha_{r,p} = \frac{\sum_{i=1}^{I \in p} \mathbb{1}(r_i == r)}{I \in p}$$

where $\mathbb{1}(r_i == r)$ is an indicator function for whether the individual's assigned race r_i is of the racial category r in $\alpha_{r,p}$. PM is useful if you are required to assign each individual a single race. One limitation of PM is that the process discards information about the uncertainty of racial classification – for example, one person with a 99.9% predicted probability of being white and another person with a 50.1% of being white will both be assigned white as their race, despite the differences in uncertainty between the two classifications. The uncertainty and the errors in PM racial assignment will be compounded when aggregating individuals up to geographic units, as one would do if using PM to estimate the racial composition of precincts. For example, if 100 voters in a precinct all each had a 50.1% probability of being white, PM would estimate that $\alpha_{white,p} = 1$ (that the precinct was 100% white), even if intuitively one might think that only about half of the voters in that precinct would actually be white.

Another method of using BISG predicted race probabilities in redistricting is called the polygon-aggregated probability summed method, which we abbreviate as PSM. This method takes an average of the predicted probabilities for all individuals within a geographic unit, such as a county, precinct, or congressional district, to estimate the racial composition of that geographic unit. Using PSM, one would calculate $\alpha_{r,p}$ with:

$$\alpha_{r,p} = \frac{\sum_{i=1}^{I \in p} P_{i,r}}{I \in p}$$

Unlike PM, the individual-level errors in racial assignment are not compounded when aggregated to geographic units. If 100 voters in a precinct all each had a 50.1% of being white, PSM would estimate that $\alpha_{white,p} = 0.501$ (that the precinct was 50.1% white), in contrast to PM. Hence, PSM should be used in most redistricting contexts that require BISG estimates of race, such as constructing majority minority districts in cases where voter race data is missing or conducting racially polarized voting estimates from precinct data with missing racial composition. More generally, researchers should use PSM in population level studies, where the errors from compounding individual-level misclassifications in racial assignment become large, and rely on PM in cases where individuals must be assigned to a singular race due to context or research design.

B Data Description

Shapefiles for precincts come from the Census Bureau's Tigerlines data set of tabular blocks.[1] We use Census block data to estimate the total population of precincts for redistricting simulations. We assigned blocks to precincts conditional upon in which precinct its geographic centroid was located (Gimpel, Lee, and Kaminski 2006). We acquired North Carolina precinct data from the North Carolina State Board of Elections (NCSBE) precinct map archive[2] and the voter file from the archive

1. U.S. Census Bureau. 2010 Census Tallies of Census, Tracts, Block Groups, and Blocks. (last updated 03/26/2012), https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2010.html (accessed 10/10/2020).

2. "Precinct Maps, 2012." North Carolina State Board of Elections. (last updated 2/8/2016), https://dl.ncsbe.gov/?prefix=PrecinctMaps/ (accessed 2/17/2021).

snapshots from the NCSBE.[3] We attained a precinct map from the Open Precincts project site for Georgia.[4] We purchased the entire Georgia voter file in December of 2020 for $250 on the Georgia Secretary of State website.[5] Georgia's voter file exhibited mismatch between the precinct IDs in the voter file and those in the precinct map, so we geocoded every address using the ESRI 2013 classic geocoder suite, and overlaid ensuing point shapefile onto the precinct shapefile.

C Diagnosing BISG

In order to diagnose the precision of the BISG estimates at the individual level, we display density plots of the uncertainty in BISG estimates in Figure S2, separately for both North Carolina (a) and Georgia (b). On the x-axis, we plot the effective number of races, the inverse of the Herfindahl index (Wolak 2009; Curiel and Steelman 2020). Figure S2 demonstrates that in most cases the BISG estimates range between one and two effective races. Overall, there is a global mode at approximately one estimated racial grouping; however, there are local modes at around two effective races, suggesting a substantive level of uncertainty in the estimation of racial categorization. Figure S3 and Figure S4 plot the average classification error rates separately for white and Black voters based on the effective number of races, in North Carolina and Georgia, respectively, and confirm that as BISG uncertainty increases so does the average error.

References

Clark, J. T., J. A. Curiel, and T. Steelman. 2021. "Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race." *Political Analysis,* 1–7.

Curiel, J. A., and T. Steelman. 2020. "A Response to "Tests for Unconstitutional Partisan Gerrymandering in a Post-Gill World" in a Post-Rucho World." *Election Law Journal* 20 (1).

Gimpel, J. G., F. E. Lee, and J. Kaminski. 2006. "The Political Geography of Campaign Contributions in American Politics." *Journal of Politics* 68:626–39.

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

Wolak, J. 2009. "The Consequences of Concurrent Campaigns for Citizen Knowledge of Congressional Candidates." *Political Behavior* 31 (2): 211–29.

3. "Voter registration snapshots." North Carolina State Board of Elections, November 6, 2012. (last updated 3/2/2017). https://dl.ncsbe.gov/index.html?prefix=data/Snapshots/ (accessed 2/17/2021).
4. Georgia. Open Precincts. <https://openprecincts.org/ga/> (accessed 2/22/2021).
5. Voter list. Georgia Secretary of State. https://georgiasecretaryofstate.net/collections/voter-list-1 (accessed 12/1/2020).

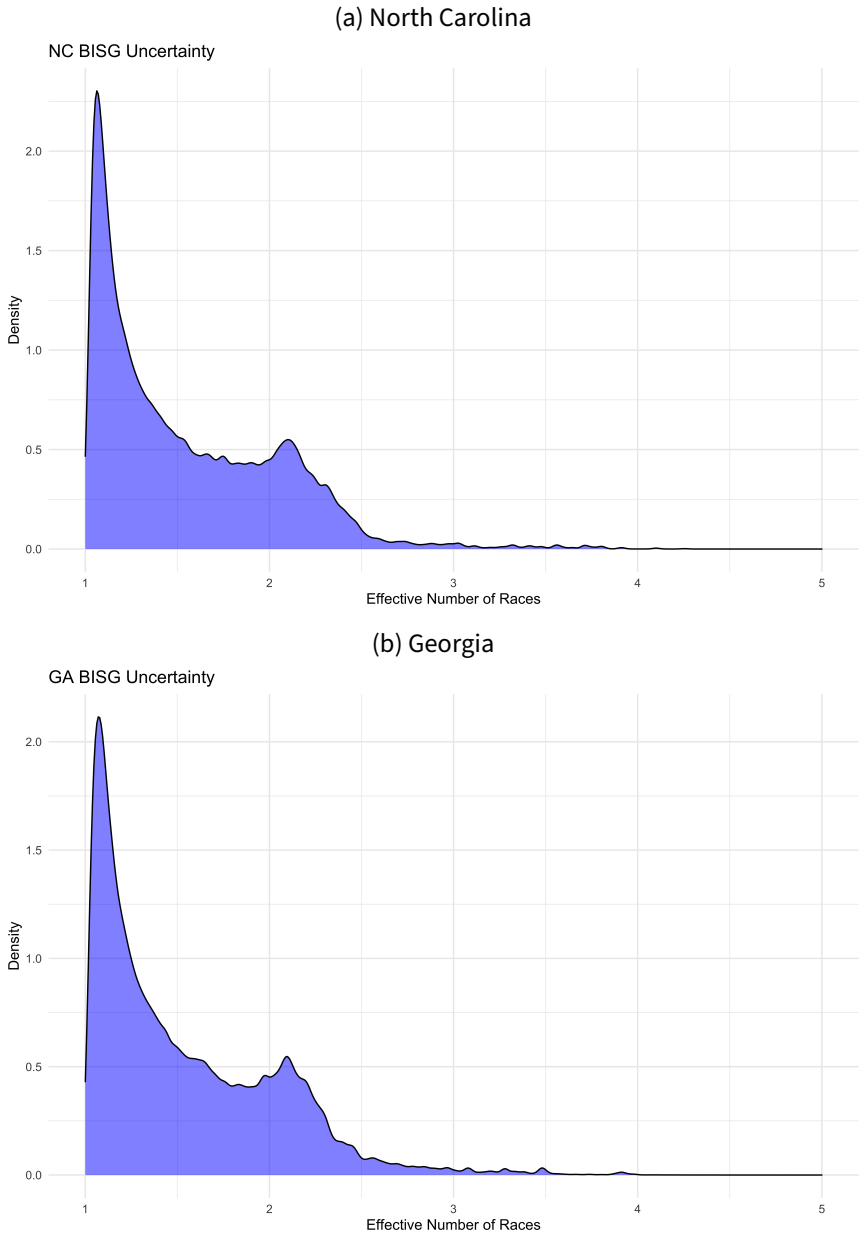

Figure S2. Precision of Individual-Level BISG Estimates

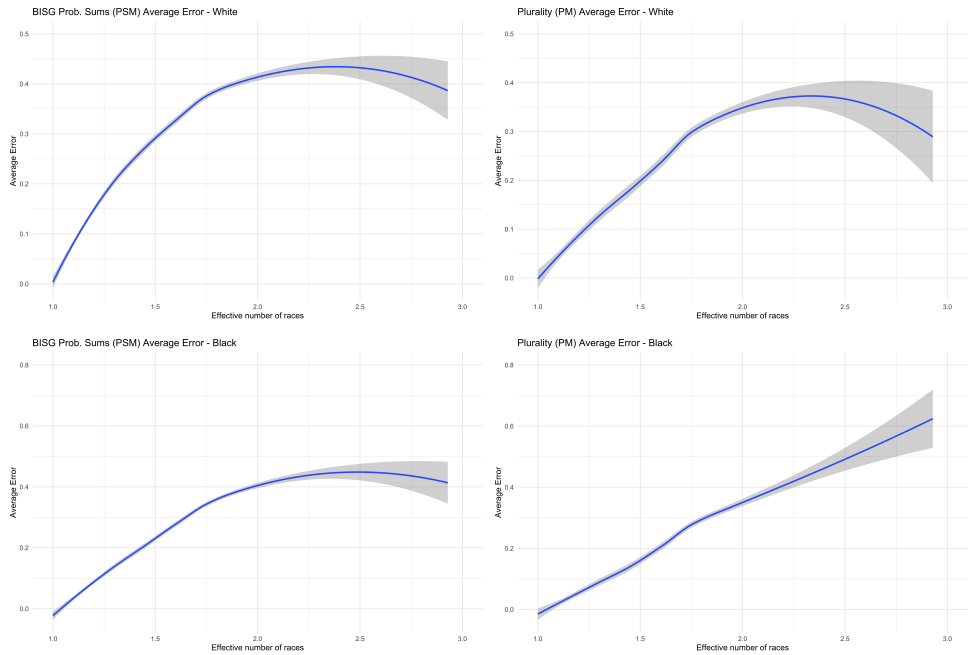

Figure S3. Errors by Heterogeneity - North Carolina

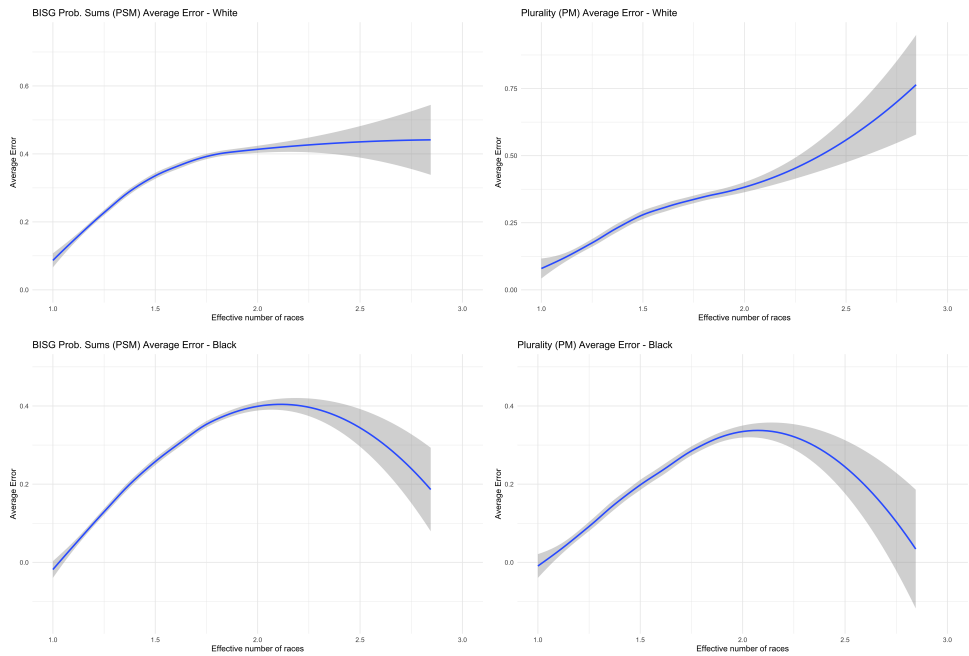

Figure S4. Errors by Heterogeneity - Georgia