PA

Validating the Applicability of Bayesian Inference with Surname and Geocoding to Congressional Redistricting

Kevin DeLuca[1] **and John A. Curiel**[2]

[1] *Ph.D. Candidate in Political Economy and Government, Harvard Kennedy School, 79 John F. Kennedy Street, Cambridge, MA 02138, USA. E-mail: kevindeluca@g.harvard.edu*
[2] *Assistant Professor of Political Science, Ohio Northern University, Hill Memorial 204B, Ada, OH 45810, USA. E-mail: j-curiel@onu.edu*

Abstract

Ensuring descriptive representation of racial minorities without packing minorities too heavily into districts is a perpetual difficulty, especially in states lacking voter file race data. One advance since the 2010 redistricting cycle is the advent of Bayesian Improved Surname Geocoding (BISG), which greatly improves upon previous ecological inference methods in identifying voter race. In this article, we test the viability of employing BISG to redistricting under two posterior allocation methods for race assignment: plurality versus probabilistic. We validate these methods through 10,000 redistricting simulations of North Carolina and Georgia's congressional districts and compare BISG estimates to actual voter file racial data. We find that probabilistic summing of the BISG posteriors significantly reduces error rates at the precinct and district level relative to plurality racial assignment, and therefore should be the preferred method when using BISG for redistricting. Our results suggest that BISG can aid in the construction of majority-minority districts during the redistricting process.

Keywords: redistricting, race, representation, Bayesian estimation, BISG

1 Introduction

The creation of majority-minority districts for underrepresented racial minorities remains a key point of contention within the field of redistricting and representation. There are the constant dangers of "packing" racial minorities into too few districts and minimizing their influence within the legislature, or "cracking" racial minorities into districts with no representatives of the same race. Striking the correct balance is not only of great normative and theoretical concern, but also methodological. It is difficult to identify the optimal racial composition of districts that avoids wasting the votes of racial minorities.

Typically, to address this problem, researchers have turned to ecological inference (EI) methodology to estimate both the electoral turnout by race and electoral preference of those who turn out given the demographics of an area (Goodman 1953; King 1997). This allows mapmakers to calculate the racial composition needed in districts to allow minorities to elect their candidate of choice. Using Census data alone, however, does not account for differential levels of registration or turnout rates across racial groups, both of which can significantly affect racial minorities' political influence.[1] Even the best EI methods coupled with on-the-ground qualitative research are far from perfect, and often necessarily lead to district plans "erring" on the side of packing minorities into districts so as to avoid unintentional cracking (Hicks *et al.* 2018).

In some cases, race data on voters contained in states' voter registration files can aid in the creation of majority-minority districts. Voter registration files contain the set of registered voters,

1 Recent evidence suggests that mapmakers target differential voter eligibility and turnout when gerrymandering (Fraga 2015; Henderson, Sekhon, and Titiunik 2016).

Political Analysis (2023)
vol. 31: 465–471
DOI: 10.1017/pan.2022.14

Published
20 May 2022

Corresponding author
Kevin DeLuca

Edited by
Jeff Gill

and often individual-level voter history. Therefore, it is possible to estimate an individual-level likely turnout model by race using a voter registration file, replacing or supplementing EI produced turnout estimates used to estimate vote choice (King 1997, 92–94). However, many states do not collect individual race data in their voter files—including states like Texas, Pennsylvania, and Wisconsin, where there is often contentious redistricting litigation.

A new development to imputing missing voter race data is Bayesian Improved Surname Geocoding (BISG) estimation. Implemented first in the field of public health by Elliott *et al.* (2008), BISG calculates the joint probability of racial membership given surname and geographic residence. Imai and Khanna (2016) find the joint information of surname and residence reduces the bias and errors of estimated race and turnout beyond that of even advanced EI methods by a magnitude of 10 (270). Therefore, the relatively new BISG methodology strictly dominates the turnout-stage estimates present within EI, which is already employed within redistricting and associated litigation. However, to date, there has been no research validating the extent to which BISG can be used to construct accurate estimates of the racial composition of proposed legislative districts.

In this letter, we estimate and proffer the best practices and baseline uncertainty for BISG in the context of redistricting. We first implement BISG on voter file data using two posterior allocation methods: polygon-aggregated probability summed method (PSM) estimates versus individual-level plurality method (PM) assignment. We show that at the precinct level, using the PSM method results in significantly lower error rates relative to PM, particularly when estimating the Black share of precinct populations, as required by the Voting Rights Act (VRA) within North Carolina and Georgia. We then estimate district-level uncertainty around each BISG method by simulating 10,000 congressional district plans for each state and compare BISG estimates of the racial composition of districts to actual voter file data. Using PSM, BISG district-level estimates of the share of minority voters in districts typically fall within five percentage points of self-reported voter file racial data, although the magnitude of the errors vary across states and racial groups. These findings demonstrate that summing probabilities produce better precinct- and district-level racial composition estimates relative to plurality assignment.

2 Using BISG in Redistricting

BISG uses an individual's surname and location to estimate their race via Bayes' rule (Elliott *et al.* 2008; Imai and Khanna 2016). Using individuals' surnames matched to a surname dictionary as the prior, joined to Census geography demographics for the conditional probability, produces more accurate racial estimates relative to other methods (Imai and Khanna 2016). While the errors tend to be greatest where surnames are uninformative and geographic units heterogeneous by race (Imai and Khanna 2016; King 1997), BISG greatly reduces the number of individuals afflicted by such uncertainty. As long as subcounty units are employed as the geography, BISG racial estimates outperform alternative methods when verified against states with race in their voter files (Clark, Curiel, and Steelman 2021; Imai and Khanna 2016). The benefits of BISG therefore earned its widespread use within political science, such as estimating the race of political donors (Alvarez, Katz, and Kim 2020; Grumbach, Sahn, and Staszak 2020), candidate emergence (Conroy and Green 2020), and minority candidate performance (Shah and Davis 2017). These developments in BISG—not available at the time of the 2010 redistricting cycle—offer an opportunity to efficiently incorporate voter race information to evaluate majority-minority districts in the current cycle.

One pressing question before applying BISG to redistricting is ascertaining the degree of error given *how* the researcher assigns racial categories from BISG posterior probability estimates. Clark *et al.* (2021) follow the practice of summing the estimated probabilities that an individual is of a given race up to a geographic unit of interest, such as a precinct. However, some scholars, such

as Enos, Kaufman, and Sands (2019), assign a single race to a voter given the racial category with the highest estimated probability, also known as deterministic, modal, or pluralistic assignment. The work by Enos *et al.* (2019) avoids substantial error by relying on segregated Los Angeles (with extremely homogeneous precincts by race), and also by dropping observations where the predicted posterior for the plurality race is under 90%. Crabtree and Chykina (2018), Rhinehart and Geras (2020), Lu *et al.* (2019), Abott and Magazinnik (2020), and Grumbach *et al.* (2020) all employ pluralistic assignment of BISG race estimates.

Plurality assignment goes against best practices within population-level social sciences (King 1997) given the potential for extreme and clustered errors. Normally, plurality assignment of race would not be considered for redistricting. In the aforementioned studies, scholars required whole assignment of their observations to a single racial group due to their research design, or relied upon statistical packages that defaulted to such assignment (Lu *et al.* 2019, 465). Plurality assignment might also appeal to redistricting practitioners; knowing each individual voters' race could allow for more sophisticated voter targeting while redistricting. However, the utility of either method depends on their accuracy when applied to redistricting. Therefore, it is important to assess the relative accuracy of both the plurality assignment and summing probabilities in the redistricting context. For a more technical explanation of BISG and the differences between PSM and PM, see Section A of the Supplementary Material.

3 BISG Validation and Simulation Results

We validate the application of BISG using two states with racial information in their voter files: North Carolina and Georgia. These states also require Black majority minority districts at the congressional level. We implemented BISG using the R package **zipWRUext** (Clark *et al.* 2021), which uses surname and ZIP code demographics to calculate the joint probability of race for individuals. While not as accurate as using addresses matched to Census block data for non-Black racial minority estimates, **zipWRUext** allows us to quickly produce accurate estimates of the predicted race of each voter using ZIP codes without having to undergo a costly and time-consuming geocoding process.[2] Section B of the Supplementary Material describes our data in more detail, and we perform diagnostics on the individual-level race BISG predictions in Section C of the Supplementary Material.

Next, we estimate the proportion of Black and White voters in each precinct, using both the PSM and the PM assignment procedures.[3] These estimates are benchmarked to the actual self-reported racial data within the voter files. Figure 1 shows a density plot of the precinct-level errors in racial estimates, calculated as the absolute percentage difference between the BISG estimated and true reported number of voters of each race. We plot the results separately for both North Carolina and Georgia. For White voters, the modal error approaches zero for both BISG assignment methods, although PM has a longer right tail, indicating worse performance relative to PSM. For Black voters, PSM vastly outperforms PM in reducing precinct-level errors. A recommendation we make confidently from just these precinct-level results is that PSM should be the preferred method when estimating the racial composition of precincts by using BISG on voter files.

To evaluate the accuracy of BISG estimates of race at the district level, we perform 10,000 redistricting simulations, each of North Carolina and Georgia's congressional district maps using

2 Geocoding millions of addresses can take weeks and cost thousands of dollars, which often presents an obstacle to utilizing BISG for those without large research budgets. In contrast, when using ZIP codes (available in all voter files that contain the addresses that would be necessary for geocoding) as the BISG geography, it takes about 10 minutes to produce racial estimates for 7 million voters in the Georgia voter file, on a 3.1-GHz MacBook Pro with 8GB of RAM.

3 We specifically analyze Black share of registered voters, given that the VRA as implemented to North Carolina and Georgia requires Blacks comprise the majority, or at least a plurality, of the district's population. Including other races in our analysis would effectively reduce our analysis to White versus non-White categories.

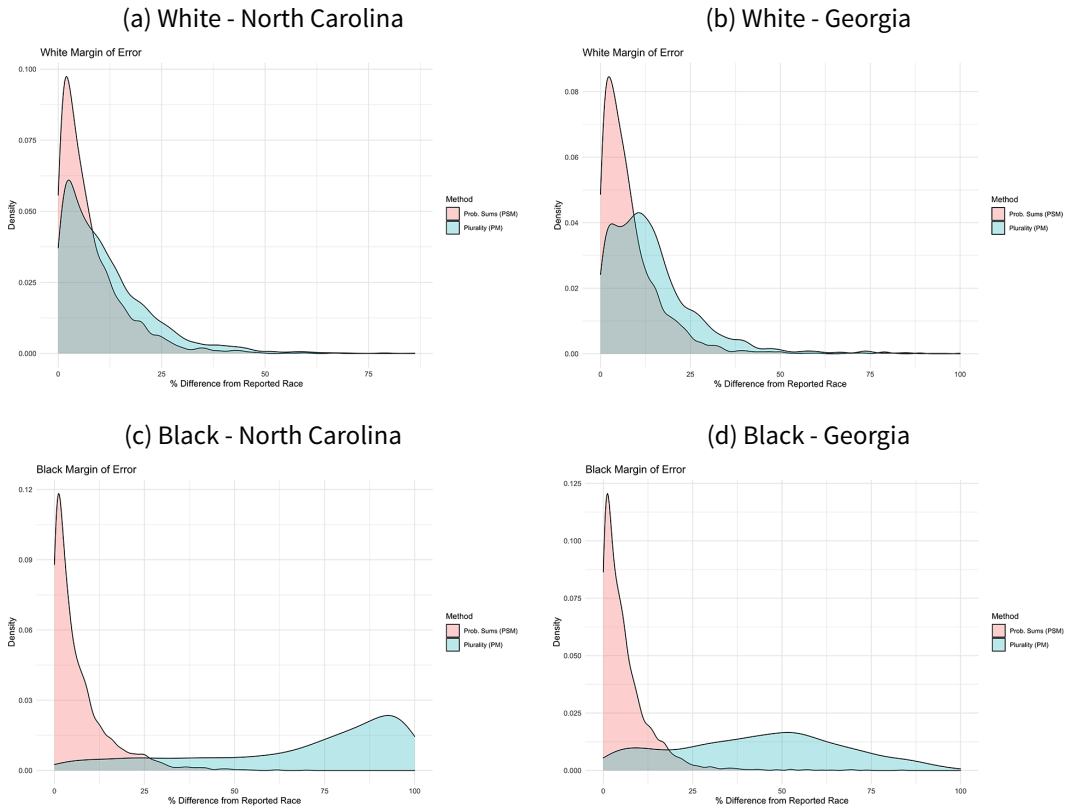

Figure 1. BISG precinct-level error density plots.

the **Redist** package in R (Fifield *et al.* 2020b), version 2. We craft a base map from the precinct simplified map employed by Curiel and Steelman (2018) for North Carolina, and do the same using Georgia's precinct shapefile for their 2010 congressional districts.[4] We then proceed to simulate districts via rook contiguity. For each simulated plan, we calculate the absolute percentage point difference between the BISG estimated proportion and the actual voter file proportion of each race in each district. We use our simulations to observe a distribution of district-level errors and create a 95% confidence interval around these estimates.

The error rates and confidence intervals for North Carolina and Georgia are plotted in Figure 2, for both White and Black voters. The *x*-axis is the share of the district population that is White when plotting the absolute error for White voters (plots (a) and (b)), and the share of the district population that is Black when plotting the absolute error for Black voters (plots (c) and (d)). In nearly all district-level estimates, summing the estimate probabilities (PSM) results in significantly lower absolute error rates relative to the PM, consistent with the precinct-level diagnostics.

While the error rates for PSM are low in general, they vary both across states and across racial groups. In North Carolina, the errors for PSM are close to zero for the percentage of White voters in each district, and never go above five percentage points for the percentage of Black voters in each district. In Georgia, the PSM error rates are slightly higher—for White voters, they max out around 10 percentage points, but for Black voters, the error rates are lower and, like North Carolina, peak around 5 percentage points. These simulation results further suggest that using the BISG PSM rather than the PM of race assignment in redistricting work will produce more accurate estimates of the racial composition of districts.

4 To successfully run simulations using **Redist**, we corrected slight errors in Georgia's shapefile. Specifically, we removed tiny holes between adjacent precinct boundaries and ensured that precinct boundaries did not overlap.

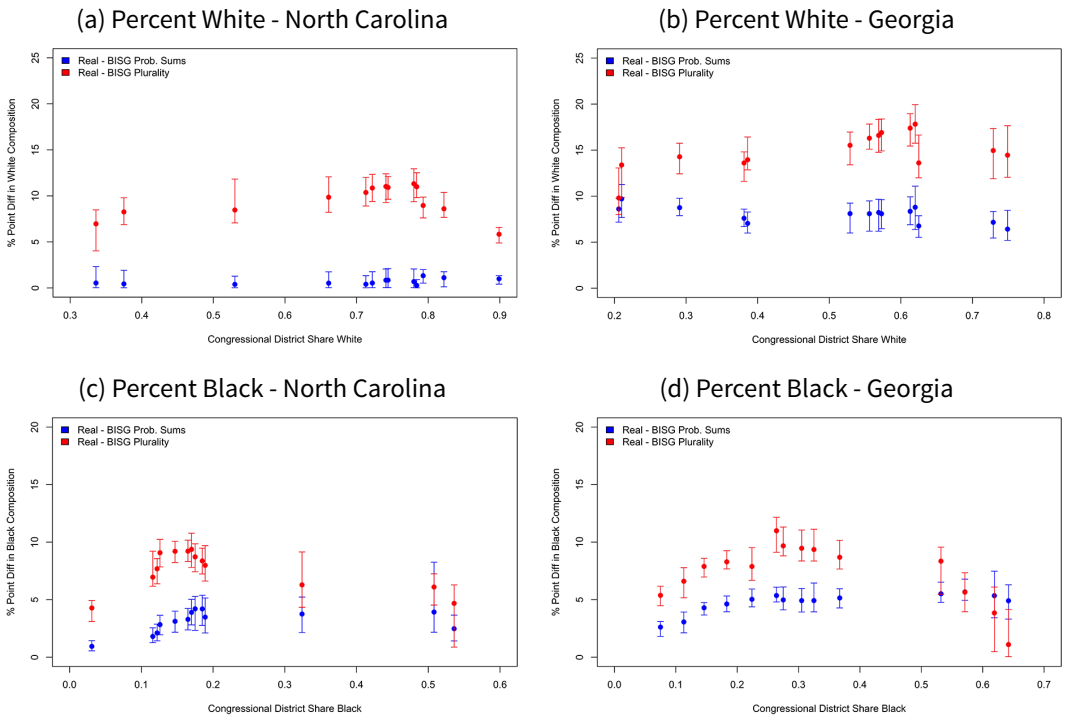

Figure 2. BISG district-level error sensitivity—North Carolina and Georgia.

4 Discussion

As simulations become more common in redistricting (Fifield *et al.* 2020a), and as the new redistricting cycle progresses without the previous protections of VRA preclearance, BISG has the potential to help provide researchers constructing optimal majority-minority districts. This can be especially useful in states where voter race data are missing from voter files. Our letter performs the first empirical validation of applying BISG in redistricting, and provides a set of simple recommendations and guidelines for researchers using BISG in redistricting analyses.

First, researchers using BISG should aggregate up to some polygonal unit of interest by summing the estimated probabilities of racial membership. Although it might be tempting to assign race to single voters in order to aid in point-based redistricting attempts, the errors will be drastically higher. Second, researchers should be prepared to deal with around a 5–10 percentage point error rate in estimating race at the district level.

In states where voter race is not collected, BISG offers a quick and fairly accurate work-around. However, the electoral context matters; insofar as electoral preferences can be divided between White and non-White categories, such as the drawing of coalition districts, BISG reaches high levels of accuracy. However, when researchers need to estimate the district composition of a specific racial minority group, such as Blacks or Hispanics, the potential for greater error should be considered.[5]

Lastly, we show that it is possible to achieve these BISG estimates at a relatively low cost via modern BISG packages in programs such as R. Imai and Khanna (2016) greatly expanded the ease of integrating Census data and surname dictionaries for BISG, and Clark *et al.* (2021) demonstrate the ability to attain accurate estimates using ZIP codes while avoiding the need to geocode altogether. We use these new methods here to provide accurate race estimates for millions of

5 Additionally, treating Hispanic as an ethnicity as opposed to race might be necessary where Hispanic-majority districts are required. Note that this would require using a package besides WRU, which treats Hispanic as a mutually exclusive race.

voters in just a couple of minutes, and we demonstrate heterogeneity in errors associated with BISG that researchers should be aware of.

Future work should look at the accuracy of BISG and redistricting in states where non-Black racial minorities and Hispanic voters make up a more significant share of the electorate. Because the two biggest racial categories in North Carolina and Georgia are Black and White, and due to significant residential segregation in each state, BISG will produce more accurate estimates in these states relative to states with more heterogeneity in racial group demographics. Other work can and should try to incorporate BISG estimates with differential turnout across racial groups from voter history (which is often contained in voter files) to create majority-minority districts.

Legislators, researchers, and everyday citizens will have access to a whole new set of quantitative tools during the 2020 redistricting cycle. Many of these tools and methods are aimed at reducing partisan and racial biases in maps to promote more fair and equal representation. However, these tools and methods can still produce biased or inefficient districts if voter race data themselves are unrepresentative of the actual electorate. Our letter helps to reduce the errors in estimating aggregate racial data, and can assist mapmakers using these new quantitative tools create efficient majority-minority districts.

Acknowledgment

We would like to thank the three anonymous reviewers and the Editor (Jeff Gill) for their thoughtful comments and discussion.

Data Availability Statement

Replication materials can be found on Harvard dataverse at Curiel and DeLuca (2022). For privacy reasons, personal identifying information from the voter files is redacted from the replication materials.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.14.

References

Abott, C., and A. Magazinnik. 2020. "At-Large Elections and Minority Representation in Local Government." *American Journal of Political Science* 64 (3): 717–733.

Alvarez, R. M., J. N. Katz, and S.-y. S. Kim. 2020. "Hidden Donors: The Censoring Problem in U.S. Federal Campaign Finance Data." *Election Law Journal* 19 (1): 1–18.

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. https://doi.org/10.1017/pan.2021.31

Conroy, M., and J. Green. 2020. "It Takes a Motive: Communal and Agentic Articulated Interest and Candidate Emergence." *Political Research Quarterly* 73 (4): 942–956.

Crabtree, C., and V. Chykina. 2018. "Last Name Selection in Audit Studies." *Sociological Science* 5: 21–28. https://doi.org/10.15195/v5.a2

Curiel, J. A., and K. DeLuca. 2022. "Replication Data for: Validating the Applicability of Bayesian Inference with Surname and Geocoding to Congressional Redistricting." Harvard Dataverse, https://doi.org/10.7910/DVN/LXGDWZ

Curiel, J. A., and T. S. Steelman. 2018. "Redistricting Out Representation: Democratic Harms in Splitting Zip Codes." *Election Law Journal* 17 (2): 328–353.

Elliott, M. N., A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie. 2008. "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." *Health Services Research* 43 (5p1): 1722–1736. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6773.2008.00854.x (accessed October 26, 2020).

Enos, R. D., A. R. Kaufman, and M. 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): 1012–1028.

Fifield, B., K. Imai, J. Kawahara, and C. T. Kenny. 2020a. "The Essential Role of Empirical Validation in Legislative Redistricting Simulation." *Statistics and Public Policy* 7 (1): 52–68.
Fifield, B., C. T. Kenny, C. McCartan, A. Tarr, and K. Imai. 2020b. "redist: Simulation Methods for Legislative Redistricting." Available at the Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=redist.
Fraga, B. L. 2015. "Redistricting and the Causal Impact of Race on Voter Turnout." *Journal of Politics* 78 (1): 19–34.
Goodman, L. 1953. "Ecological Regressions and the Behavior of Individuals." *American Sociological Review* 18: 663–666.
Grumbach, J., A. Sahn, and S. Staszak. 2020. " *Gender, Race, and Intersectionality in Campaign Finance*." *Political Behavior* 44: 319–340.
Henderson, J. A., J. S. Sekhon, and R. Titiunik. 2016. "Cause or Effect? Turnout in Hispanic Majority-Minority Districts." *Political Analysis* 24 (3): 404–412.
Hicks, W. D., C. E. Klarner, S. C. McKee, and D. A. Smith. 2018. "Revisiting Majority-Minority Districts and Black Representation." *Political Research Quarterly* 71 (2): 408–423.
Imai, K., and K. Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Record." *Political Analysis* 24 (2): 263–272.
King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.
Lu, C., et al. 2019. "Examining Scientific Writing Styles from the Perspective of Linguistic Complexity." *Journal of the Association for Information Science and Technology* 70 (5): 462–475.
Rhinehart, S., and M. J. Geras. 2020. "Diversity and Power: Selection Method and Its Impacts on State Executive Descriptive Representation." *State Politics and Policy Quarterly* 20 (2): 213–233.
Shah, P. R., and N. R. Davis. 2017. "Comparing Three Methods of Measuring Race/Ethnicity." *Journal of Race, Ethnicity, and Politics* 2 (1): 124–139.