PA

Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race

Jesse T. Clark[1], **John A. Curiel**[2] and **Tyler S. Steelman**[3]

[1] *Postdoctoral Research Associate, Princeton University, Princeton, NJ*
[2] *Assistant Professor, Ohio Northern University, Ada, OH, USA*
[3] *Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.*
E-mail: tsteelman@unc.edu

Abstract

Racial identification is a critical factor in understanding a multitude of important outcomes in many fields. However, inferring an individual's race from ecological data is prone to bias and error. This process was only recently improved via Bayesian improved surname geocoding (BISG). With surname and geographic-based demographic data, it is possible to more accurately estimate individual racial identification than ever before. However, the level of geography used in this process varies widely. Whereas some existing work makes use of geocoding to place individuals in precise census blocks, a substantial portion either skips geocoding altogether or relies on estimation using surname or county-level analyses. Presently, the trade-offs of such variation are unknown. In this letter, we quantify those trade-offs through a validation of BISG on Georgia's voter file using both geocoded and nongeocoded processes and introduce a new level of geography—ZIP codes—to this method. We find that when estimating the racial identification of White and Black voters, nongeocoded ZIP code-based estimates are acceptable alternatives. However, census blocks provide the most accurate estimations when imputing racial identification for Asian and Hispanic voters. Our results document the most efficient means to sequentially conduct BISG analysis to maximize racial identification estimation while simultaneously minimizing data missingness and bias.

Keywords: Bayesian improved surname geocoding, racial identification, geocoding, geographic information system, ZIP codes

1 Introduction

Within American politics, racial identification is a key variable across a wide range of sub-fields, though individual measures of racial identification are not always readily available. To address the lack of data on racial identification, past research first turned to ecological inference (see king 1997; Robinson 1950) before marked improvements in the form of "Bayesian improved surname geocoding" (BISG) to use individual names and locations to impute the race of a single individual (Elliott *et al.* 2008).

In their 2016 *Political Analysis* article, Imai and Khanna (2016) extended the utility of BISG by increasing its ease of use to a wider field of scholarship via their R package **wru** while simultaneously reducing imputation error and bias in estimating an individual's racial identification. Their framework has been applied to a variety of research questions where individual-level racial data are difficult to obtain (Alvarez, Katz, and Kim 2020; Edwards, Esposito, and Lee 2018; Signorella 2020). Table 1 highlights that variety and further demonstrates the range of geocoding and nongeocoding methods used to implement BISG.[1]

Political Analysis (2022)
vol. 30: 456–462
DOI: 10.1017/pan.2021.31

Published
29 November 2021

Corresponding author
Tyler S. Steelman

Edited by
Jeff Gill

1 Overall, the plurality of work within political science estimates BISG with geocoded data, with half of these making use of census blocks to estimate local level turnout (Fraga 2018), local policy changes following riots (Enos, Kaufman, and Sands 2019), and even presence of hidden donors on presidential campaigns (Alvarez *et al.* 2020). Fields outside of political

Table 1. Overview of Bayesian improved surname geocoding levels employed in previous articles, by field.

Field	Surname only	County	Geocoded	Unclear
Political Science	28.6% (4)	21.4% (3)	42.9% (6)	7.1% (1)
Sociology	40.0% (2)	40.0% (2)	20% (1)	0
Public Health	25.0% (1)	50% (2)	25% (1)	0
Other	33.3% (1)	33.3% (1)	33.3% (1)	0

While imputing race is more accessible than ever, there are two notable limitations that remain unaddressed. First, the trade-offs in accuracy when employing varying levels of geography to estimate racial identification using **wru** are unclear. Second, geocoding—which involves turning address data into coordinates (Swift, Goldberg, and Wilson 2008) can be costly in both time and resources and results in substantial missing data (Amos and McDonald 2020).[2] Barriers to estimation arising from geocoding costs are most notable when using the voter file format employed by Imai and Khanna (2016) at the census tract level and below.

We address these issues by extending **wru** in two ways. First, we introduce a new level of geography in the form of ZIP codes as a means to conduct BISG without geocoding. Second, we conduct BISG for every level of geography available in **wru**—and ZIP codes—to test the accuracy of racial estimates using various geographic levels. We test the accuracy of each level using the Georgia voter file, one of seven states that includes the self-reported race of registered voters, by examining accuracy differences by race for a combination of geocoded (tracts and blocks) and nongeocoded geographic units (surname alone, county, and ZIP code). Overall, we find that the degree of accuracy in racial identification estimation is effectively indistinguishable between census tracts and ZIP codes, with ZIP codes emerging the preferred alternative given its reduction of data missingness and its avoidance of geocoding altogether. However, when estimating Hispanic, Asian, and "other" races,[3] census block level estimates are preferred.

This letter contributes to existing work utilizing BISG in two ways. First, we clarify the trade-offs of using various levels of geography—like Census blocks, Census tracts, and counties—when estimating racial identification using BISG. Secondly, we extend the accessibility of BISG to researchers through the introduction of ZIP codes in BISG estimation. The evidence presented here demonstrates that ZIP codes, as the smallest unit of publicly known geography, can meet existing levels of accuracy without the added need of geocoding. This allows researchers to simplify the estimation of racial identification by using nongeocoding methods first and more costly geocoding methods only when necessary.

2 Validating Different Levels of Geography

To clarify the trade-offs in accuracy when estimating racial identification using BISG using a variety of geographies, we predict the probability an individual is of a given race using five approaches. Specifically, we conduct a BISG analysis using the nongeocoded methods of surname only, county, and ZIP code and geocoded methods using census tract and census blocks. To test the accuracy of each method, we predict the race of each voter in the Georgia voter file and validate our prediction using the self-reported racial identification data found in the same source. The Georgia voter file contains 7,346,219 records, with 3,123,112 unique addresses (Clark, Curiel, and Steelman 2021).

science are far less likely to employ geocoded data in their BISG predictions, instead relying on surname only or county-level analyses (Einstein, Glick, and Palmer 2020; Lu *et al.* 2019; Studdert *et al.* 2020).

2 Missing data vary by geocoders, with more accurate geocoders usually being proprietary. Furthermore, data missingness often correlate with race (Swift *et al.* 2008). In addition, costs can soar among the best geocoders. The Google API, for example, has a rate of $5.00/1,000 addresses. See, Geocoding API Usage and Billing. Google Maps Platform. https://developers.google.com/maps/documentation/geocoding/usage-and-billing (accessed June 1, 2020).

3 Categorized as not falling into the groups of White, Black, Hispanic, or Asian/Pacific Islander.

We introduce ZIP codes to BISG analyses given that they are the smallest unit of publicly known geography and therefore do not need to be geocoded for use. There are approximately 30,000 ZIP codes in the United States. While there is some variance in the population of individual ZIP codes, the overall population distribution is strikingly similar to the population distribution of census tracts (Curiel and Steelman 2018). The Georgia voter file includes a ZIP code for each registered voter in the state. Although it is not uncommon for states to also include each voter's county, counties have a much higher degree of population variance and significantly higher levels of racial segregation when compared to ZIP codes (Nall 2018; Nemerever and Rogers 2021).

For the purposes of our analysis, we create an R package called **zipWRUext**, which takes the framework of **wru** and supplements the existing structure to work with ZIP code level census and American Community Survey (ACS) data. This extension calculates the joint probability of racial identification given the ZIP code and surname of an individual. Furthermore, this package allows users to specify a given year from 2010 to 2018 with either census or ACS data to improve the imputation of an individual's racial identification to align with the user's research context. For the purpose of this analysis, we employ estimates for ZIP codes using 2010 census and 2018 ACS data. Missing data were restricted to the 2,065 registered voters for whom no ZIP code was available and represent 0.03% of the data.

The package **wru** includes two geocoding alternatives: census tracts and blocks. Regardless of which geographic unit is used, both incur the same standard costs in time and money to geocode. Amos and McDonald (2020) demonstrate the most recent advances in proper geocoding for spatial audit purposes and find that even powerful computers can take several hours to complete a full geocoding process and will still result in a missing rate of 1–3 percentage points (6). Swift *et al*. (2008) report that the five best geocoders at the time of writing—including ESRI—feature a missing rate of around 5%.

Although geocoded data can produce more accurate imputations of racial identification than nongeocoded alternatives, users continue to incur costs related to geocoding and must eventually use nongeocoded data to locate individuals that cannot be geocoded. In this letter, we made use of the ESRI 2013 street address and postal address geocoders to locate Georgian addresses which took several hours to complete on 3,123,112 unique addresses.[4] This method was unable to geocode 4.6% of addresses. In the end, the Georgia voter file placed voters in 3,113 unique census tracts and 156,301 census blocks. The two computationally demanding processes are the geocoding of addresses and the overlaying of those addresses onto census geographic data, followed by the importing of the census demographics at the block level. Overall, the geocoding process took 7.28 hr, as specified in Tables 3 and 4 in the Supplementary Material.

Using these geocoded and nongeocoded data, we use the **wru** package to impute the racial identification for all Georgian voters into the categories of White, Black, Hispanic, Asian, and others.[5] We then bootstrap 10,000 draws, with samples of 1,000 in each draw, to determine the accuracy of each method relative to self-reported racial identification. We store the actual number for each race drawn in addition to the sum of the race probability estimates for each of the BISG level estimates. Finally, we calculate the absolute difference between these BISG estimates to report the distributional difference both as percent differences. This allows us to create a distribution of uncertainty as opposed to a static state-level population estimate.

4 The computer specifications are Intel(R) core i7-7500U 2.90 GHz, 16-GB RAM.

5 Georgia, like Florida, treats Hispanic as a race as opposed to an ethnicity. Therefore, Georgia works well for validation given the categories present within the voter file unlike a state like North Carolina, which treats and codes Hispanic as an ethnicity.

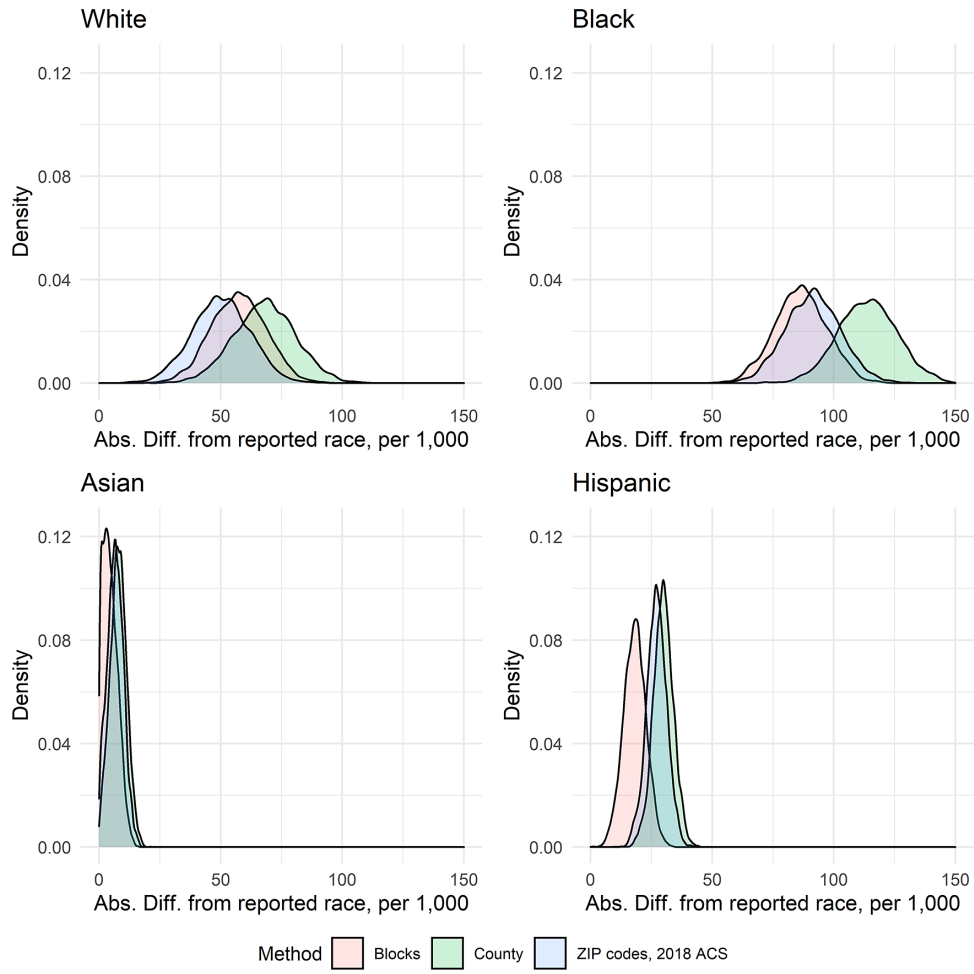

Figure 1. Count difference between reported and estimated race for blocks, ZIP codes, and counties by race.

3 Results

We report the detailed results of the 10,000 bootstraps in Table 3 in the Supplementary Material. Overall, the results are such that ZIP codes, tracts, and blocks produce the most accurate results on the whole. Estimates for ZIP codes and census tracts are virtually identical, with exceptions being marginal and not substantive when imputing White and Black racial identification. In addition, census blocks are the most useful geocoded geographic unit when imputing Asian and Hispanic racial identification.

Figure 1 visualizes the differences in accuracy using the absolute difference in count out of 1,000 from the reported race drawn from the 10,000 bootstraps for blocks, ZIP codes, and counties. By race, we see that the three plotted methods are highly clustered when estimating White racial identification, although counties are the least accurate. Blocks and ZIP codes reach similar levels of accuracy when estimating Black racial identification with counties again underperforming. We also find that ZIP codes outperform counties when estimating Asian and Hispanic identification, although blocks continue to be more accurate in both cases. The estimates are such that blocks and ZIP codes are substantively the same for estimating White and Black racial identification within the bootstrapped data.

ZIP codes and census blocks produce very similar levels of accuracy even though ZIP codes do not require complex geocoding. The median difference between ZIP codes and blocks for White and Blacks are −1.22 and 1.54 per 1,000 draws, respectively. However, the difference in medians between ZIP codes and blocks reaches 9.21 and 23.35 per 1,000 draws for Asian and Hispanic racial identification, respectively. As a comparison to the next level of precision used

in BISG, tracts, the median difference between ZIP codes and tracts for White and Black racial identification are −0.15 and 0.12 per 1,000 draws, respectively. For Asian and Hispanic racial identification, the difference in medians between ZIP codes and tracts amounts to 0.05 and 4.81 per 1,000 draws, respectively. Therefore, blocks, tracts, and ZIP codes are effectively equivalent in accuracy for estimating White and Black populations. For Asian and Hispanic populations, block estimates exceed that of both tracts and ZIP codes, with the latter two being substantively equivalent.

Given the cost in geocoding and processing times, which took 7.28 hr to complete in this analysis, the gain in reducing the median difference in error per hour are 1.27 and 3.21 for Asian and Hispanic racial identification, respectively. Such gains can be seen as substantive and worth the added investment of geocoding. For White and Black racial identification, the rates are −0.17 and 0.21. The gains for tracts over ZIP codes, in turn, are effectively zero despite the additional hours necessary to geocode.

4 Discussion of Applications

Our findings provide a path forward for the use of BISG in estimating racial group identification in a variety of new applications. Through our analysis of the Georgia voter file, we are able to confirm that the **wru** implementation of Imai and Khanna (2016) continues to perform exceedingly well, and the extension provided here will further its applicability. This letter can serve as a guide for future researchers trying to distinguish between where geocoding is necessary and where nongeocoding alternatives—like ZIP codes—can be employed without sacrificing accuracy.

It is important to note two potential challenges that can characterize this type of analysis. First, modifiable areal unit problems (MAUPs) can introduce noise into any analysis that relies on geographic shape files (Duque, Laniado, and Polo 2018). As a result, the accuracy findings presented in this letter can be considered more conservative estimates. Furthermore, the noise introduced by MAUPs in the case of Georgia may not apply in the same way to other geographic contexts. In addition, racial identification as a fluid concept can introduce error when using self-reported racial identification, especially among Asian and Hispanic individuals (Masuoka 2006; Masuoka, Ramanathan, and Junn 2019). Special attention should be paid when computing racial identification to ensure that imputations are a reflection of the lived experience of the subjects being studied and account for the trichotomy of race, ethnicity, and nationality (Masuoka 2006).

Future work should increase the transparency of how analysts implement BISG when discussing their geocoding process, rationale, and geographic units. While we were able to identify the primary level of BISG geocoding in most articles citing Imai and Khanna (2016), it was often unclear what geographic unit was used and how missing data were handled. As noted in this letter, the trade-offs of using geocoded versus nongeocoded imputation methods come with significant costs to data missingness and accuracy. Transparency must be a central tenant of any research utilizing these methods.

ZIP codes are the superior alternative to other nongeocoded BISG processes. Furthermore, ZIP codes are effectively on par to estimates derived using census tracts without the added costs associated with geocoding. Should a researcher find themselves in a situation where geocoding their data is either necessary or preferred, census blocks should be used as opposed to census tracts given their higher accuracy when estimating racial identification. County and surname only estimates are error prone and should only be used when no other alternative is viable. However, we recognize that context matters; in cases where researchers are only estimating the racial identification of White and Black individuals, ZIP code estimates are effectively indistinguishable

from block-level estimates while allowing researchers to avoid geocoding and spatial overlap costs.[6]

We recommend that researchers incorporate ZIP codes into future BISG research. Researchers should only use more computationally costly geocoded alternatives when required; surname only and county-level analyses should only be used when all other alternatives have been examined. Such sequential BISG predictions can robustly reduce estimation error when imputing the racial identification of individuals.

Acknowledgments

We would like to thank the two anonymous reviewers and the Editor, Jeff Gill, for their thoughtful comments and discussion.

Data Availability Statement

The replication materials for this paper can be found on the Harvard Dataverse at Clark *et al.* (2021). For privacy reasons, personal identifying information is redacted from the replication materials.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2021.31.

References

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

Amos, B., and M. P. McDonald. 2020. "A Method to Audit the Assignment of Registered Voters to Districts and Precincts." *Political Analysis* 28(3):356–371.

Clark, J., J. A. Curiel, and T. S. Steelman. 2021. "Replication Data for: Minmaxing of Bayesian Improved Surname and Geography Level Ups in Predicting Race." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/IH7ICK.

Curiel, J. A., and T. S. Steelman. 2018. "Redistricting out Representation: Democratic Harms in Splitting Zip Codes." *Election Law Journal* 17(4):328–353.

Duque, J. C., H. Laniado, and A. Polo. 2018. "S-Maup: Statistical Test to Measure the Sensitivity to the Modifiable Areal Unit Problem." *PLoS One* 13(11):1–25.

Edwards, F., M. H. Esposito, and H. Lee. 2018. "Risk of Police-Involved Death by Race/Ethnicity and Place, United States, 2012–2018." *American Journal of Public Health* 108(9):1241–1248.

Einstein, K. L., D. M. Glick, and M. Palmer. 2020. *Neighborhood Defenders: Participatory Politics and America's Housing Crisis*. Cambridge: Cambridge University Press.

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://doi.org/10.1111/j.1475-6773.2008.00854.x.

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.

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

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: 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.

Masuoka, N. 2006. "Together they Become One: Examining the Predictors of Panethnic Group Consciousness Among Asian Americans and Latinos." *Social Science Quarterly* 87(5):993–1011.

6 Loading in census block data is computationally expensive, often requiring several gigabytes of space. The spatial overlay process added 20 min. While not a major issue for Georgia, it can be prohibitive for some scholars in larger states, such as New York, Florida, Texas, and California.

Masuoka, N., K. Ramanathan, and J. Junn. 2019. "New Asian American Voters: Political Incorporation and Participation in 2016." *Political Research Quarterly* 72(4):991–1003.
Nall, C. 2018. *The Road to Inequality: How the Federal Highway Program Polarized America and Undermined Cities*. Cambridge: Cambridge University Press.
Nemerever, Z., and M. Rogers. 2021. "Measuring the Rural Continuum in Political Science." *Political Analysis* 29(3):1–20.
Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15(3):351–357. https://doi.org/10.2307/2087176.
Signorella, M. L. 2020. "Toward a More Just Feminism." *Psychology of Women Quarterly* 44(2):256–265. https://doi.org/10.1177/0361684320908320.
Studdert, D. M., et al. 2020. "Handgun Ownership and Suicide in California." *New England Journal of Medicine* 382(23):2220–2229.
Swift, J. N., D. W. Goldberg, and J. P. Wilson. 2008. "Geocoding Best Practices: Review of Eight Commonly Used Geocoding Systems." Technical report 10, University of Southern California Research GIS Laboratory, Los Angeles. https://spatial.usc.edu/wp-content/uploads/2014/03/gislabtr10.pdf.