Supplementary Materials for "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]*Department of Political Science, Massachusetts Institute of Technology.*
[2]*Department of Political Science, Massachusetts Institute of Technology.*
[3]*Department of Political Science, University of North Carolina-Chapel Hill. Email: tsteelman@unc.edu*

Previous BISG Work

Table 1. Sources citing Imai and Khanna (2016) and Prediction Level

Source	Field	Level
Abott and Magazinnik (2020)	Political Science	County
Alvarez, Katz, and Kim (2020)	Political Science	Blocks
Conroy and Green (2020)	Political Science	Surname
Enos, Kaufman, and Sands (2019)	Political Science	Blocks
Fraga (2018)	Political Science	Blocks
Grumbach and Sahn (2020)	Political Science	Tract
Grumbach, Sahn, and Staszak (2020)	Political Science	Tracts
Hood III, Morrison, and Bryan (2018)	Political Science	Precincts/blocks
Reny, Wilcox-Archuleta, and Nichols (2018)	Political Science	Unclear
Rhinehart and Geras (2020)	Political Science	Surname
Sadhwani and Mendez (2018)	Political Science	Surname
Schwemmer and Jungkunz (2019)	Political Science	Surname
Shah and Davis (2017)	Political Science	County
Velez and Newman (2019)	Political Science	County
Barreto *et al.* (2019)	Sociology	Surname
Chou, Imai, and Rosenfeld (2020)	Sociology	County
Crabtree and Chykina (2018)	Sociology	County
Einstein, Glick, and Palmer (2020)	Sociology	Tracts
Signorella (2020)	Sociology	Surname
Edwards, Esposito, and Lee (2018)	Public Health	Counties
Nguyen *et al.* (2019)	Public Health	County
Riester *et al.* (2019)	Public Health	Surname
Studdert *et al.* (2020)	Public Health	Tract
Edwards, Lee, and Esposito (2019)	Other	County
Grinberg *et al.* (2019)	Other	Surname
Lu *et al.* (2019)	Other	Geocoded/not clear

Political Analysis (2021)

DOI: 10.1017/pan.xxxx.xx

Corresponding author
Tyler S. Steelman

Edited by
John Doe

Geocoding results

Figure 1 provides the screenshot output of the ESRI 2013 composite geocoding, which comprises the street address and postal geocoders. With the 3,123,112 unique addresses within the voter file,

Table 2. Addresses Geocoded Within the Georgia Voter File, by Geocoder

Geocoder	% Geocoded of Voter Addresses
Street addresses	84.71%
Postal	15.19%
N/A	0.10%

Table 3. Geocoding Locator Percentage of GA Voter file sample

Locator	Number	Percent
PointAddress	262588	84.08
StreetAddress	44083	14.11
Postal	3068	0.98
StreetName	2410	0.77
AdminPlaces	166	0.05

The proportion geocoded by ESRI locator. Total time taken amounted to geocode was 131 minutes and 10 seconds, a rate of approximately 2,384.084 per minute.

the total time taken was 4 hours and 45 minutes using an Intel(R) core i7-7500U 2.90 GHz, 16 GB RAM computer.

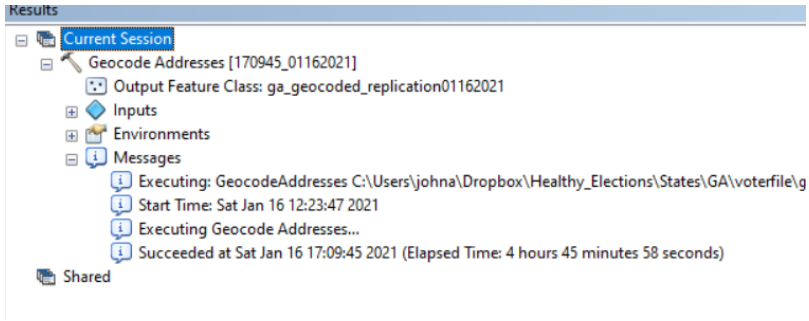

Figure 1. Screen capture of geocoding times of the Georiga voter file comprising 3,123,112 unique addresses

Table 2 reports the proportion of addresses geocoded by the ESRI 2013 Street Address geocoder, Postal geocoder, and those not geocoded. The proportions are such that 86.01 percent of addresses within the Georgia voter file can be geocoded using the street addresses geocoder. The postal address geocoder in turn provides just under 14 percent of coordinates as a back up to the priority street address geocoder. Finally, under a tenth of a percentage point of addresses cannot be geocoded.

Above is the output of the geocoded addresses used for the analysis. It is possible to increase the matching accuracy via higher level proprietary geocoders. Advances in geocoding with special licenses from ESRI do allow for more precise estimates. Therefore, we additionally ran the 2019 classic locator suite of geocoders from ESRI, which includes: Administrative Places, Point Address, Postal, PostalExtension, Street Address, Street name, and ZIP-4 locators. The suite works such that it iterates through these locators in order to identify the best coordinate match for a given address. The trade off is an increase in time and computational power necessary for geocoding.

Table 3 presents the output of geocoding a 10 percent sample of the unique addresses within Georgia. The computer used to geocode had the following specifications: Intel(R) Core(TM) i7K 8700K 6-Core/12-Thread, 12MB Cache, up to 4.7GHz with Intel(R) Turbo Boost Technology), 64GB HyperX(TM) DDR4 Memory XMP at 2933MHz, 2TB M.2 PCIe x4 SSD. The computer was part of the MIT GIS Lab, made available for researchers with intensive geocoding and other GIS needs.

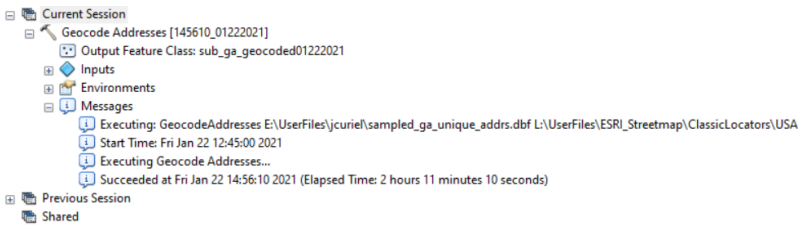

Figure 2. Screen capture of geocoding times of the 10 percent sampled Georgia voter file, comprising 312,315 unique addresses

Figure 2 documents the time taken to geocode the 10 percent sample as 2 hours, 11 minutes, and 10 seconds. The geocoding rate is approximately equal to 2,384.084 per minute. Therefore, the total time necessary to geocode the full voter file, assuming a relatively constant rate, would be around 21 hours.

Upon completing the geocoding, the next step is to overlay the addresses onto census geographies, then employ the “get_census_data" command to pull the necessary demographic information in order to conduct BISG. The computer used to run the replication file had the following specifications: Intel(R) Core(TM) i7-9700 CPU 6-Core/12-Thread, 12MB Cache, 3.0 GHz, 64GB DDR4 Memory XMP at 2933MHz, 2TB HDD. The total time necessary to overlay and get the census block and tract data amounted to 2.53 hours. Therefore, the total time taken to process and prepare the data necessary to conduct BISG dependent upon geocoding methods took approximately 7.28 hours for the Georgia voter file. Employing the full suite of proprietary ESRI geocoders would amount to over 23 hours of processing time.

BISG Accuracy estimates

We report the percent absolute difference at the 95^{th} percentile between the reported and predicted races by geographic level and race in Table 4 in the supplementary information. By race, the accuracy from most to least is in the order of White, Black, Asian, Hispanic, and other. Ranked by level of geography, surname alone and county are the least accurate. Adding in county information reduces the error for White and Black estimates, though increases it for Asian, Hispanic, and other estimates. While it is almost certainly the case that the exact errors will vary by state given their unique demographic distributions, these results are concerning for past research that opts to use counties as the primary level of geography for BISG.

Interesting results arise for ZIP codes. The ZIP code estimates using both 2010 census and 2018 ACS data consistently outperform county level data in reducing error. ZIP codes likewise outperform surname alone, with the exception of estimating racial identification for Asian and other. The reduction in error in estimating racial identification varies across racial groups: White racial identification error is reduced by 3 percentage points, Black racial identification error is reduced by 5.53 points, Asian racial identification error is reduced by 1.52 points, Hispanic racial identification error is reduced by 9.21 points, and all other racial identification error is reduced by 0.87 points. The 2018 ACS data improves upon 2010 data marginally, with the exception of estimates for Hispanic racial identification.

The added benefit of geocoding to increase accuracy in racial identification estimates arises entirely when using block-level BISG estimation. The accuracy rates for 2010 census tracts compared to 2018 ZIP codes and 2018 ACS data are effectively on par with one another.[1] Census blocks are

1. The **wru** package contains the “get_census_data" function, which uses 2010 Census data given that one has a Census API. The **zipWRUext** package pulls in the relevant ZIP code data internally as part of the base function and requires no additional Census API credential.

Table 4. Comparison of BISG accuracy by geography and race

Level	Asian	Black	Hispanic	Other	White
Surname	51.04	42.77	118.23	109.79	19.63
County	62.70	38.32	120.20	113.43	16.05
ZCTA 2010	58.38	32.79	110.99	112.56	13.02
ZCTA 2018	56.86	31.74	111.11	109.98	12.49
Tract	56.97	31.51	104.95	124.44	12.53
Block	47.57	30.23	83.74	104.98	13.40

The results are the percent absolute difference at the 95th percentile between the reported and predicted races by level and race for the 10,000 draws of samples of 1,000 records from the Georgia voter file. Larger values reflect greater errors for the given method by race.

superior in every regard to other forms of geocoded and non-geocoded estimation processes when using BISG for every racial identification except White.

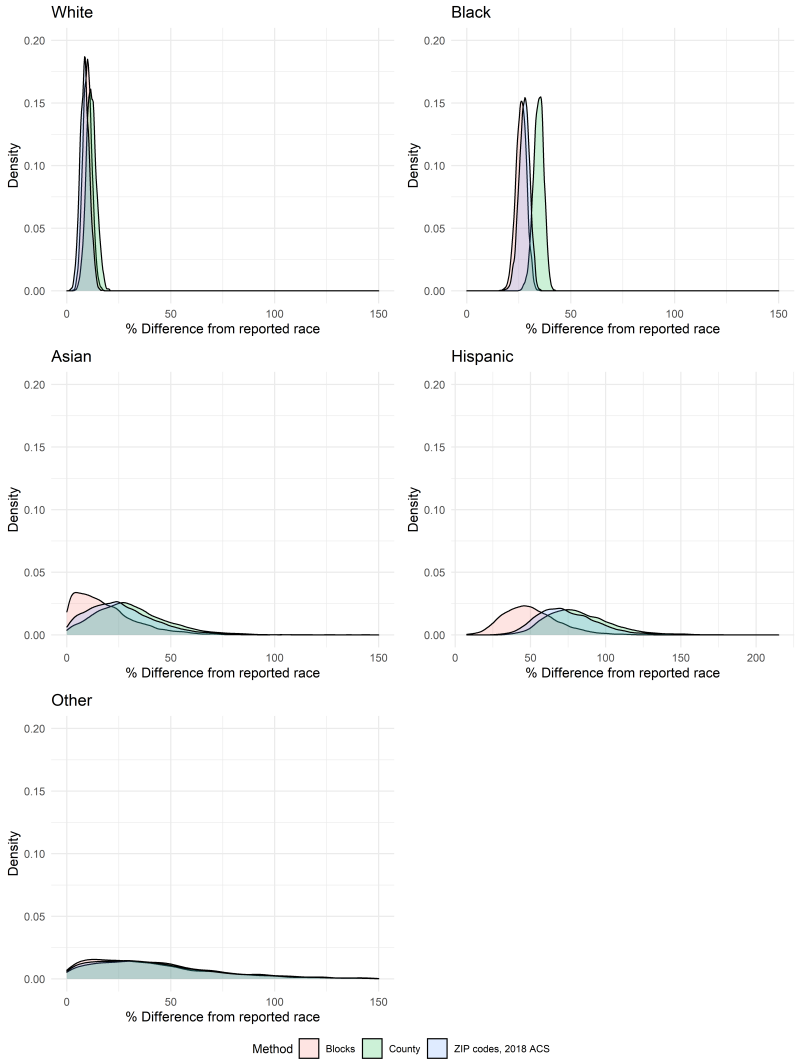

Figure 3. Percent difference between reported and estimated race for blocks, ZIP codes, and counties by race

Supplementary Material

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

References

Abott, C., and A. Magazinnik. 2020. "At-Large Elections and Minority Representation in Local Government." *American Journal of Political Science* 64 (3): 717–33.

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.

Barreto, M., L. Collingwood, S. Garcia-Rios, and K. A. Oskooii. 2019. "Estimating Candidate Support in Voting Rights Act Cases: Comparing Iterative EI and EI-R×C Methods." *Sociological Methods & Research.*

Chou, W., K. Imai, and B. Rosenfeld. 2020. "Sensitive Survey Questions with Auxiliary Information." *Sociological Methods & Research* 49 (2): 418–54.

Conroy, M., and J. Green. 2020. "It Takes a Motive: Communal and Agentic Articulated Interest and Candidate Emergence." *Political Research Quarterly* Forthcoming.

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

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): 1241–48.

Edwards, F., H. Lee, and M. 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): 16793–8.

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

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): 1012–28.

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

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. 2019. "Fake news on Twitter during the 2016 U.S. presidential election." *Science* 363 (6425): 374–8.

Grumbach, J., A. Sahn, and S. Staszak. 2020. "Gender, Race, and Intersectionality in Campaign Finance." *Political Behavior* Forthcoming.

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

Hood III, M. V., P. A. Morrison, and T. M. Bryan. 2018. "From Legal Theory to Practical Application: A How-To for Performing Vote Dilution Analyses*." *Social Science Quarterly* 99 (2): 536–552.

Lu, C., Y. Bu, J. Wang, Y. Ding, V. Torvik, M. Schnaars, and C. Zhang. 2019. "Examining scientific writing styles from the perspective of linguistic complexity." *Journal of the Association for Information Science and Technology* 70 (5): 462–475.

Nguyen, V. T., R. D. Zafonte, J. T. Chen, K. Z. Kponee-Shovein, S. Paganoni, A. Pascual-Leone, F. E. Speizer, *et al.* 2019. "Mortality Among Professional American-Style Football Players and Professional American Baseball Players." *JAMA Network Open* 2 (5).

Reny, T., B. Wilcox-Archuleta, and V. C. Nichols. 2018. "Threat, Mobilization, and Latino Voting in the 2018 Election." *The Forum* 16 (4): 573–599.

Rhinehart, S., and M. J. Geras. 2020. "Diversity and Power: Selection Method and Its Impacts on State Executive Descriptive Representation." *State Politics & Policy Quarterly* 20 (2): 213–233.

Riester, S. M., K. L. Leniek, A. D. Niece, A. Montoya-Barthelemy, W. Wilson, J. Sellman, P. J. Anderson, *et al.* 2019. "Occupational medicine clinical practice data reveal increased injury rates among Hispanic workers." *American Journal of Industrial Medicine* 62, no. 4 (April): 309–316. https://doi.org/10.1002/ajim.22949.

Sadhwani, S., and M. Mendez. 2018. "Candidate Ethnicity and Latino Voting in Co-Partisan Elections." *California Journal of Politics and Policy* 10 (2).

Schwemmer, C., and S. Jungkunz. 2019. "Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks." *Political Research Exchange* 1 (1): 1–23.

Shah, P. R., and N. R. Davis. 2017. "Comparing Three Methods of Measuring Race/Ethnicity." *Journal of Race, Ethnicity and Politics* 2 (1): 124–39.

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., Y. Zhang, S. A. Swanson, L. Prince, J. A. Rodden, E. E. Holsinger, M. J. Spittal, G. J. Wintemute, and M. Miller. 2020. “Handgun Ownership and Suicide in California.” *New England Journal of Medicine* 382 (23): 2220–9.

Velez, Y. R., and B. J. Newman. 2019. “Tuning In, Not Turning Out: Evaluating the Impact of Ethnic Television on Political Participation.” *American Journal of Political Science* 63 (4): 808–23.