A Comparative Evaluation of Proxy Estimation Methods for Racial Classification

S. Khan[1], L. Yu[1], Y. Kang[1], S. Venkatasubramanian[1]
[1] Center for Tech Responsibility (CNTR), Brown University, USA

GitHub Link:
https://github.com/brown-cntr/RaceProxyBench

BROWN
Data Science Institute

Introduction

- Bayesian Improved Surname Geocoding (BISG)[1]
  - Method for predicting race given location and surname
  - Basis for newer variants (e.g., cBISG) that patch its limitations
  - Insights into effectiveness of proxy estimation methods
- Focus: North Carolina 2022 Voter Registration Dataset[2]
- Idea: Compare proxy method outcomes by varying noise for:
  a) ZCTA/Zip Code (α) b) Surname (γ)

Methods

BISG:

$$P(R=r \mid S=s, G=g) = \frac{P(S=s \mid R=r)\, P(R=r \mid G=g)}{\sum_{r'} P(S=s \mid R=r')\, P(R=r' \mid G=g)}$$

BIFSG[3]: augments BISG with a first-name factor to improve precision for minority groups

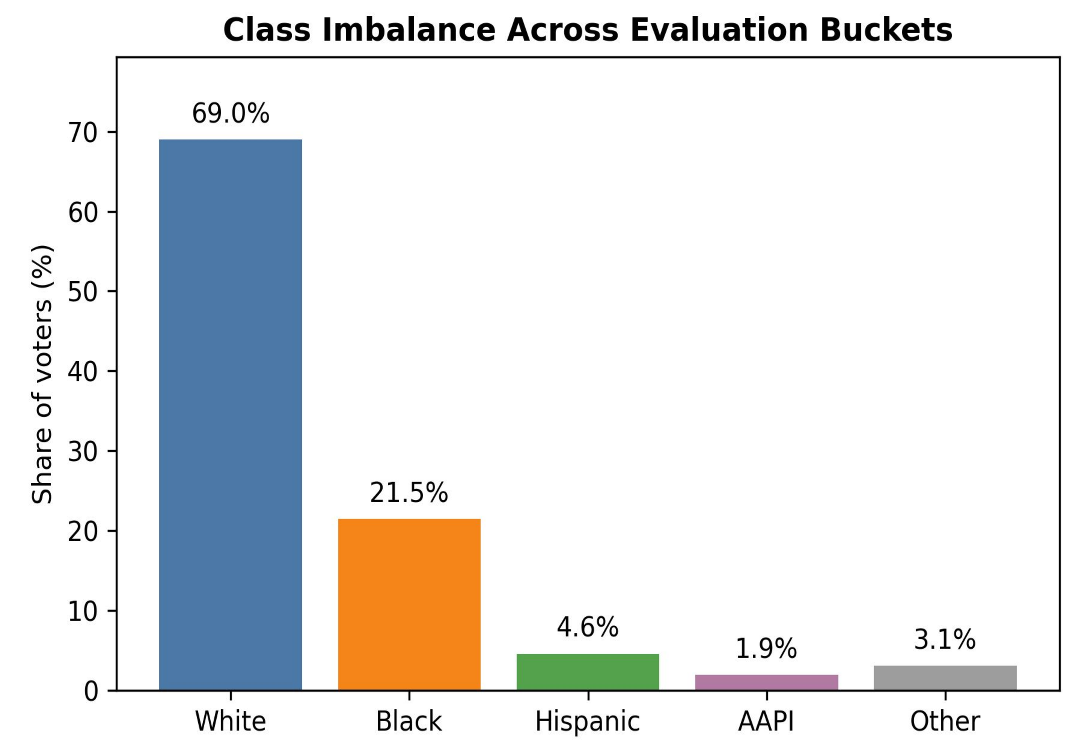
fBISG[4]: performs full posterior inference to handle surname coverage gaps and Census under-counting

cBISG[5]**:** adds contextual features (e.g. loan size, party affiliation) as extra priors

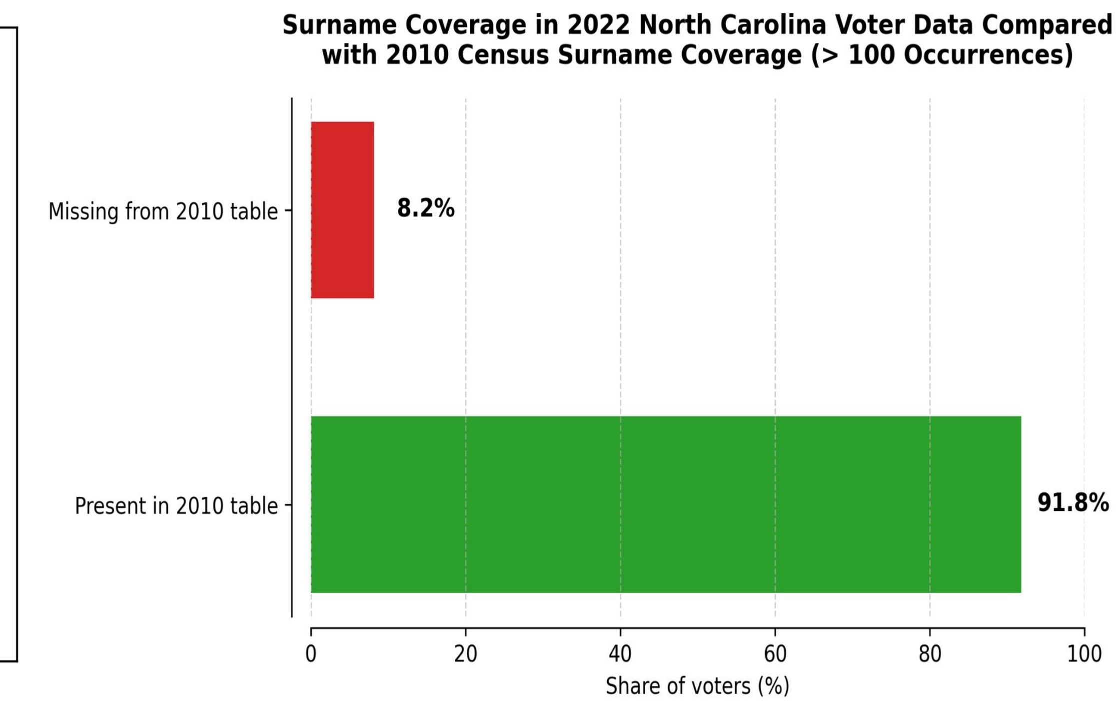
Zest Race Predictor[6]**:** trains XGBoost gradient-boosted trees on names and geographic context

Exploratory Data Analysis

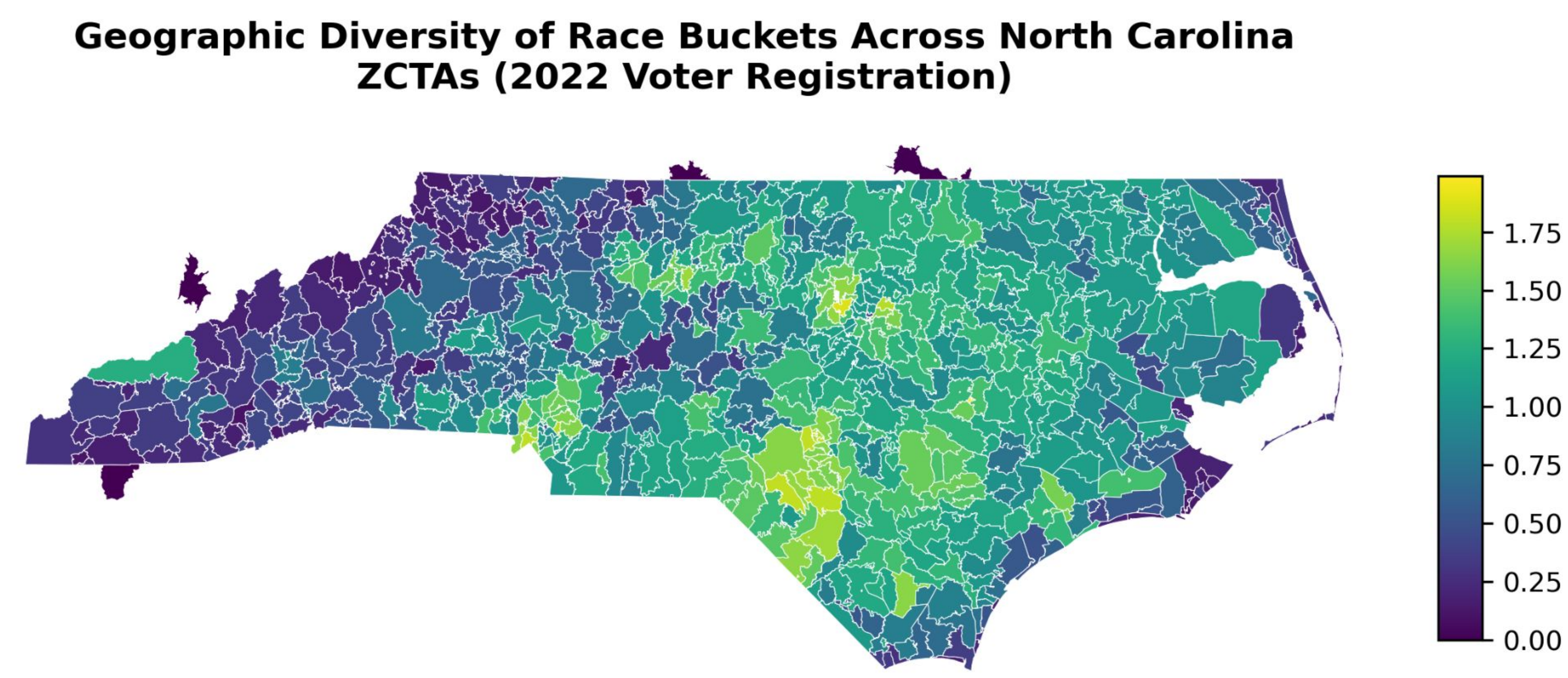
Race Mix



Name Coverage



Geo Diversity



Results

BISG

α (%)	γ (%)	Acc	F1	LL	ECE_{10}	Cov
0	0	0.7901	0.6143	2.9751	0.0190	1.0000
5	0	0.7868	0.6109	3.0031	0.0171	0.9968
10	0	0.7837	0.6077	3.0297	0.0152	0.9936
20	0	0.7770	0.6007	3.0855	0.0116	0.9871
0	5	0.7874	0.6066	2.9571	0.0217	1.0000
5	5	0.7841	0.6031	2.9852	0.0197	0.9968
10	5	0.7809	0.5998	3.0120	0.0177	0.9936
20	5	0.7742	0.5927	3.0680	0.0134	0.9871
0	10	0.7847	0.5985	2.9394	0.0245	1.0000
5	10	0.7814	0.5950	2.9676	0.0224	0.9968
10	10	0.7782	0.5917	2.9945	0.0203	0.9936
20	10	0.7714	0.5844	3.0509	0.0158	0.9871

BIFSG

α (%)	γ (%)	Acc	F1	LL	ECE_{10}	Cov
0	0	0.8341	0.6593	3.6464	0.0104	1.0000
5	0	0.8318	0.6567	3.6719	0.0097	0.9968
10	0	0.8294	0.6544	3.6961	0.0090	0.9936
20	0	0.8247	0.6492	3.7473	0.0094	0.9871
0	5	0.8323	0.6549	3.6244	0.0107	1.0000
5	5	0.8300	0.6523	3.6500	0.0100	0.9968
10	5	0.8276	0.6499	3.6744	0.0093	0.9936
20	5	0.8229	0.6447	3.7257	0.0089	0.9871
0	10	0.8305	0.6504	3.6027	0.0109	1.0000
5	10	0.8281	0.6478	3.6285	0.0102	0.9968
10	10	0.8258	0.6453	3.6529	0.0095	0.9936
20	10	0.8210	0.6401	3.7046	0.0085	0.9871

fBISG

α (%)	γ (%)	Acc	F1	LL	ECE_{10}	Cov
0	0	0.7753	0.5816	3.0659	0.0077	1.0000
5	0	0.7722	0.5786	3.0712	0.0059	0.9968
10	0	0.7692	0.5758	3.0732	0.0043	0.9936
20	0	0.7628	0.5701	3.0866	0.0036	0.9871
0	5	0.7729	0.5729	3.0434	0.0104	1.0000
5	5	0.7699	0.5699	3.0489	0.0083	0.9968
10	5	0.7668	0.5671	3.0517	0.0066	0.9936
20	5	0.7604	0.5613	3.0643	0.0036	0.9871
0	10	0.7706	0.5640	3.0213	0.0131	1.0000
5	10	0.7675	0.5608	3.0279	0.0107	0.9968
10	10	0.7644	0.5580	3.0297	0.0089	0.9936
20	10	0.7579	0.5521	3.0433	0.0051	0.9871

ZRP

α (%)	γ (%)	Acc	F1	LL	ECE_{10}	Cov
0	0	0.8564	0.6535	0.4986	0.0210	1.0000
5	0	0.8564	0.6535	0.4986	0.0210	1.0000
10	0	0.8229	0.5998	0.6229	0.0463	1.0000
20	0	0.8229	0.6223	0.5933	0.0429	1.0000
0	5	0.8185	0.6109	0.6082	0.0446	1.0000
5	5	0.8185	0.6110	0.6081	0.0446	1.0000
10	5	0.8185	0.6109	0.6082	0.0446	1.0000
20	5	0.8185	0.6109	0.6081	0.0446	1.0000
0	10	0.8140	0.5998	0.6231	0.0462	1.0000
5	10	0.8141	0.5998	0.6229	0.0463	1.0000
10	10	0.8141	0.5998	0.6229	0.0463	1.0000
20	10	0.8141	0.5998	0.6229	0.0463	1.0000

cBISG*

α (%)	γ (%)	Acc	F1	LL	ECE_{10}	Cov
0	0	0.8074	0.5749	6.7127	0.0157	1.0000
0	5	0.8050	0.5635	6.6880	0.0148	1.0000
0	10	0.8027	0.5523	6.6640	0.0140	1.0000
0	20	0.7979	0.5285	6.6151	0.0123	1.0000

**Due to cBISG's dependence on census tract for geographical context rather than ZCTA or zipcode, we applied noise to surname while maintaining the noise proportion for the geoID at zero. The input dataset for the cBISG experiments was a subset of the one used for the other four methods due to the lack of census tract availability for every sample.*

References

1. Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., & Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Services and Outcomes Research Methodology, 9(2), 69–83. https://doi.org/10.1007/s10742-009-0047-1
2. Data available at https://www.ncsbe.gov/results-data/voter-registration-data
3. Bayesian Improved First Name and Surname Geocoding (BIFSG) Voicu, I. (2018). Using first name information to improve race and ethnicity classification. Statistics and Public Policy, 5(1), 1–13. https://doi.org/10.1080/2330443X.2018.1427012
4. Imai, K., Olivella, S., & Rosenman, E. T. R. (2022). Addressing Census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements. Science Advances, 8, eadc9824. https://doi.org/10.1126/sciadv.adc9824
5. Kwegyir-Aggrey, K., Durvasula, N., Wang, J., & Venkatasubramanian, S. (2024). Observing Context Improves Disparity Estimation when Race is Unobserved (arXiv:2409.01984). https://doi.org/10.48550/arXiv.2409.01984
6. Zest AI. (2020). Zest Race Predictor (ZRP) [Computer software]. GitHub repository. https://github.com/zestai/zrp

Contributions

S. Khan - Introduction, Methods, Exploratory Data Analysis, Results (ZRP); L. Yu - Results (cBISG); Y. Kang - Results (BISG, BIFSG, fBISG); S. Venkatasubramanian - Advisor