Utilizing Bayesian Improved Surname Geocoding (BISG) to Explore California's Diverse Electorate

Michael Rios, UCLA VRP Data Scientist

VOTING RIGHTS PROJECT UCLA

About the UCLA Voting Rights Project

The UCLA Voting Rights Project is aimed at creating an accessible and equitable system of voting for all Americans through impact litigation, research, and clinical education to expand access to the ballot box.

Overview

1 THE DIVERSITY OF CALIFORNIA

2 WHAT IS BAYESIAN IMPROVED SURNAME GEOCODING (BISG)?

3 RESEARCH METHODOLOGY AND PROJECT OBJECTIVES

4 UTILIZING BISG TO EXAMINE THE CALIFORNIA ELECTORATE

How Can we Analyze Diverse Jurisdictions?

	California	
Total	26,042,367	-
White (not Hispanic or Latino)	10,673,998	41.0%
Hispanic	8,508,628	32.7%
Asian American or Pacific Islander	4,070,322	15.6%
Black	1,641,315	6.3%
All Other	1,148,104	4.4%

- California is one of the most diverse states in the U.S.
- While the U.S. Census Bureau can provide valuable data on voter eligibility, it cannot tell us about voting patterns.
- ***The problem***: Census data doesn't provide detailed voter data and the CA voter file only has self-reported race data for 24.3% of all L.A. County registered voters (about 1.4M of 5.6M).

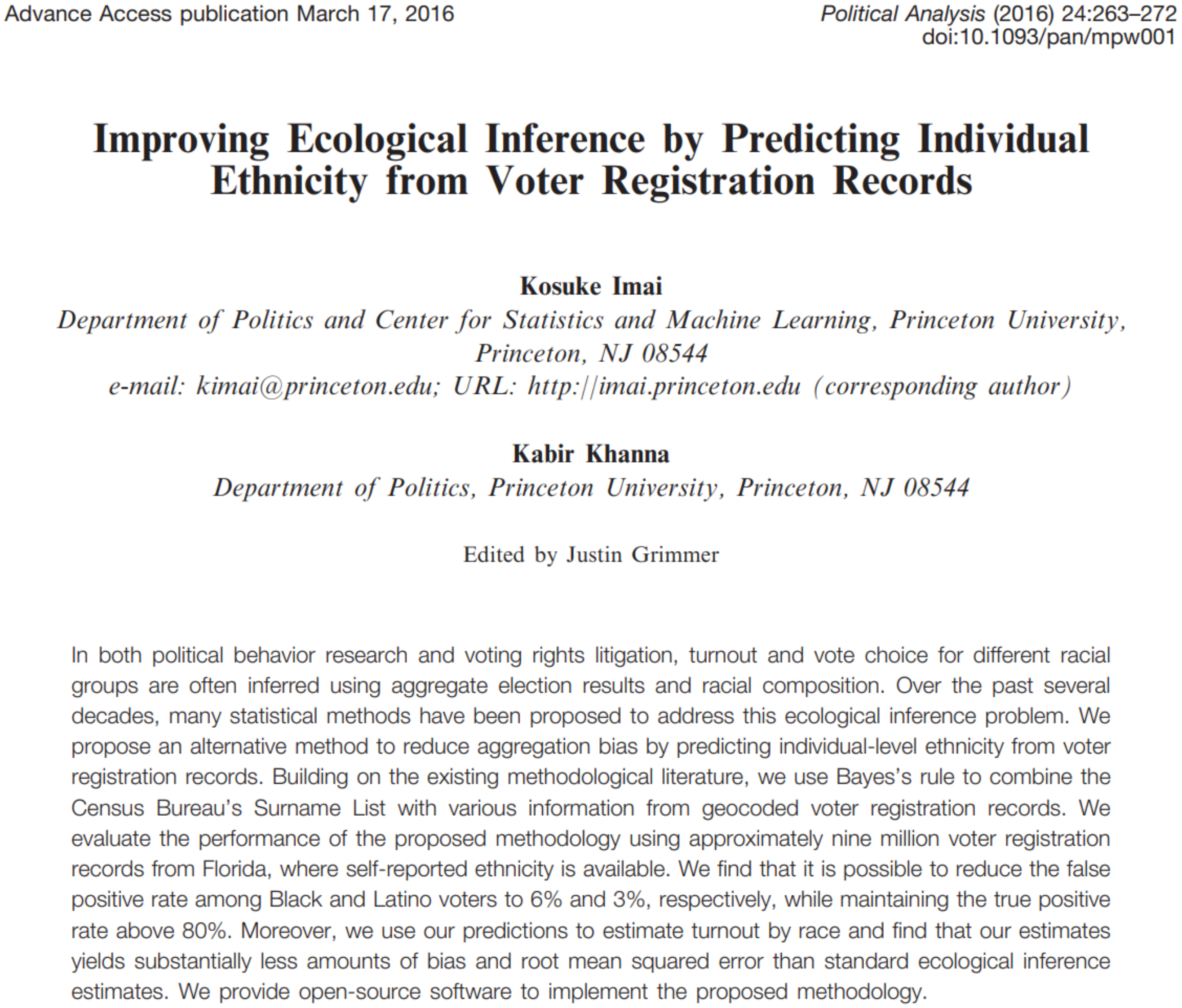
Source: U.S. Census Bureau, 2023 1-Year American Community Survey, Sex By Age By Nativity And Citizenship Status.

VOTING RIGHTS PROJECT UCLA

What is Bayesian Improved Surname Geocoding (BISG)?

Widespread use of BISG

PA
Volume 31 • Number 4 • October
POLITICAL ANALYSIS
THE SOCIETY FOR POLITICAL METHODOLOGY
CAMBRIDGE UNIVERSITY PRESS



Advance Access publication March 17, 2016

Political Analysis (2016) 24:263–272
doi:10.1093/pan/mpw001

Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records

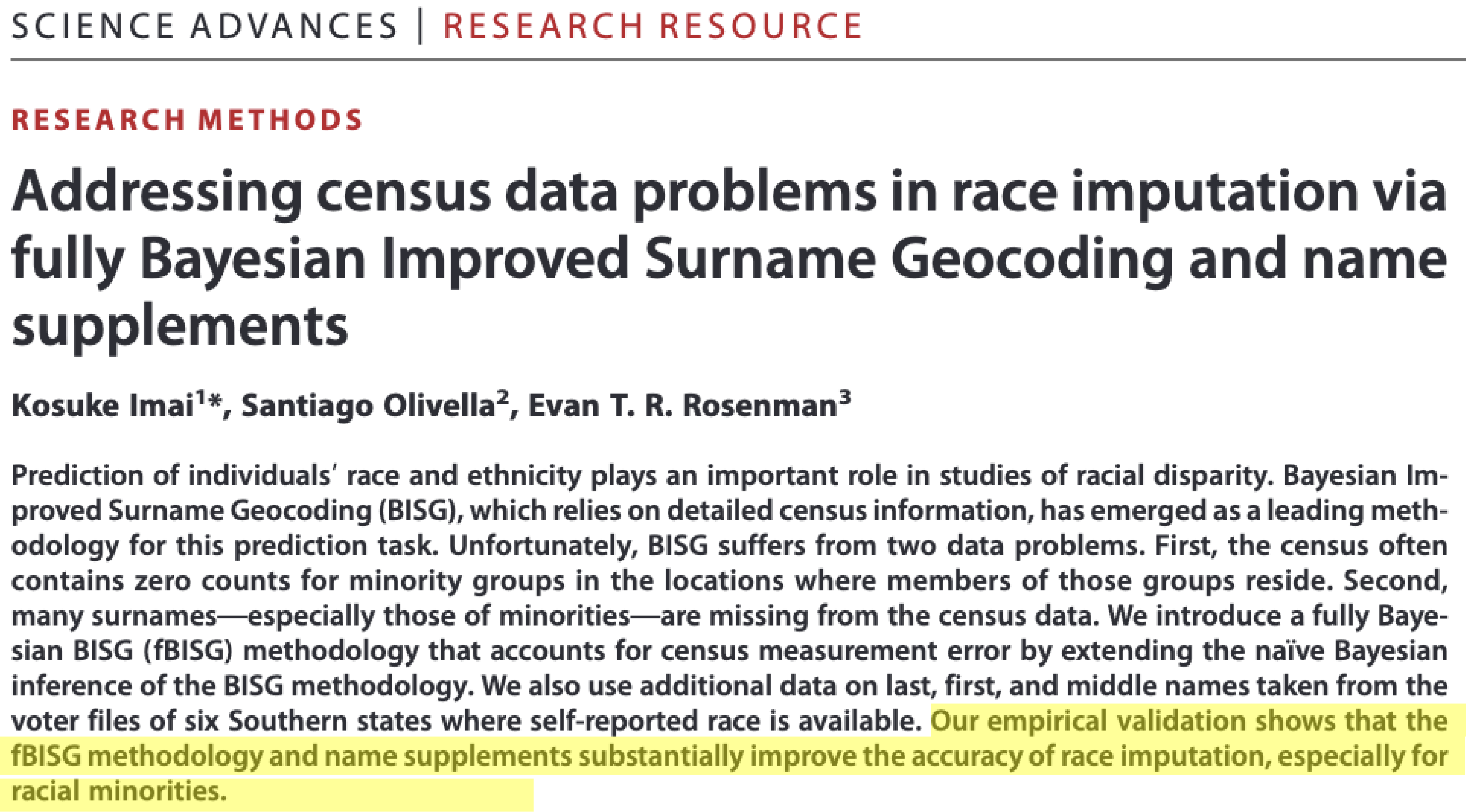
Kosuke Imai
Department of Politics and Center for Statistics and Machine Learning, Princeton University,
Princeton, NJ 08544
e-mail: kimai@princeton.edu; URL: http://imai.princeton.edu (corresponding author)

Kabir Khanna
Department of Politics, Princeton University, Princeton, NJ 08544

Edited by Justin Grimmer

In both political behavior research and voting rights litigation, turnout and vote choice for different racial groups are often inferred using aggregate election results and racial composition. Over the past several decades, many statistical methods have been proposed to address this ecological inference problem. We propose an alternative method to reduce aggregation bias by predicting individual-level ethnicity from voter registration records. Building on the existing methodological literature, we use Bayes's rule to combine the Census Bureau's Surname List with various information from geocoded voter registration records. We evaluate the performance of the proposed methodology using approximately nine million voter registration records from Florida, where self-reported ethnicity is available. We find that it is possible to reduce the false positive rate among Black and Latino voters to 6% and 3%, respectively, while maintaining the true positive rate above 80%. Moreover, we use our predictions to estimate turnout by race and find that our estimates yields substantially less amounts of bias and root mean squared error than standard ecological inference estimates. We provide open-source software to implement the proposed methodology.

Widespread use of BISG



SCIENCE ADVANCES | RESEARCH RESOURCE

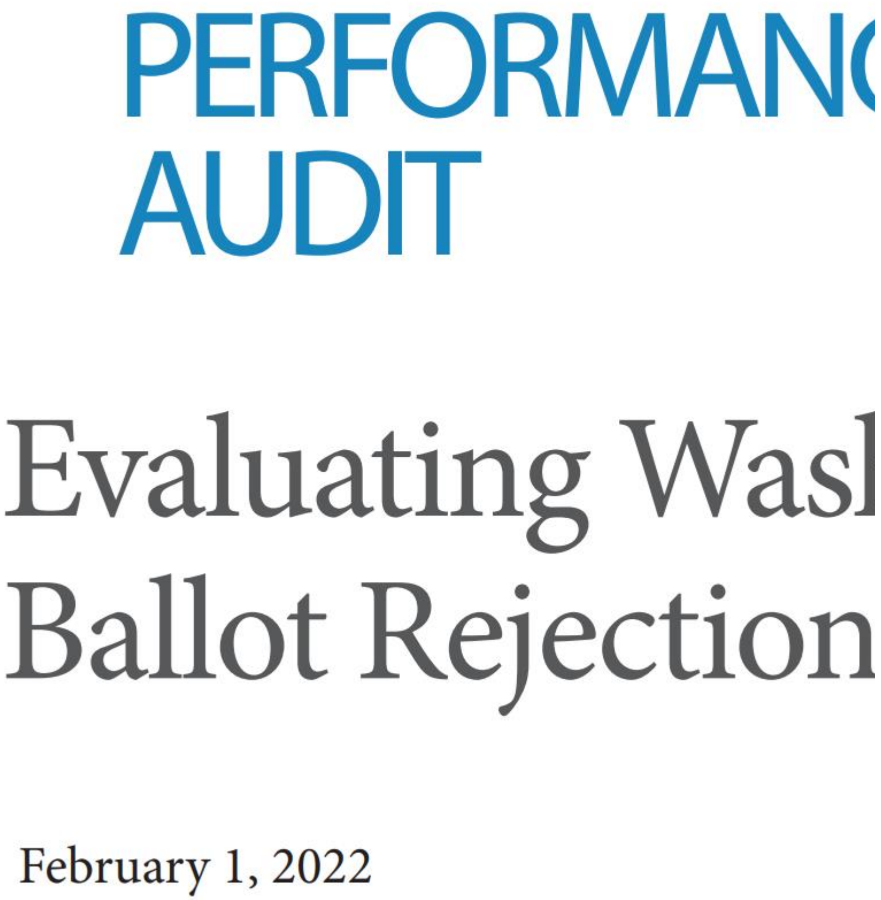
RESEARCH METHODS

Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements

Kosuke Imai[1]*, Santiago Olivella[2], Evan T. R. Rosenman[3]

Prediction of individuals' race and ethnicity plays an important role in studies of racial disparity. Bayesian Improved Surname Geocoding (BISG), which relies on detailed census information, has emerged as a leading methodology for this prediction task. Unfortunately, BISG suffers from two data problems. First, the census often contains zero counts for minority groups in the locations where members of those groups reside. Second, many surnames—especially those of minorities—are missing from the census data. We introduce a fully Bayesian BISG (fBISG) methodology that accounts for census measurement error by extending the naïve Bayesian inference of the BISG methodology. We also use additional data on last, first, and middle names taken from the voter files of six Southern states where self-reported race is available. Our empirical validation shows that the fBISG methodology and name supplements substantially improve the accuracy of race imputation, especially for racial minorities.

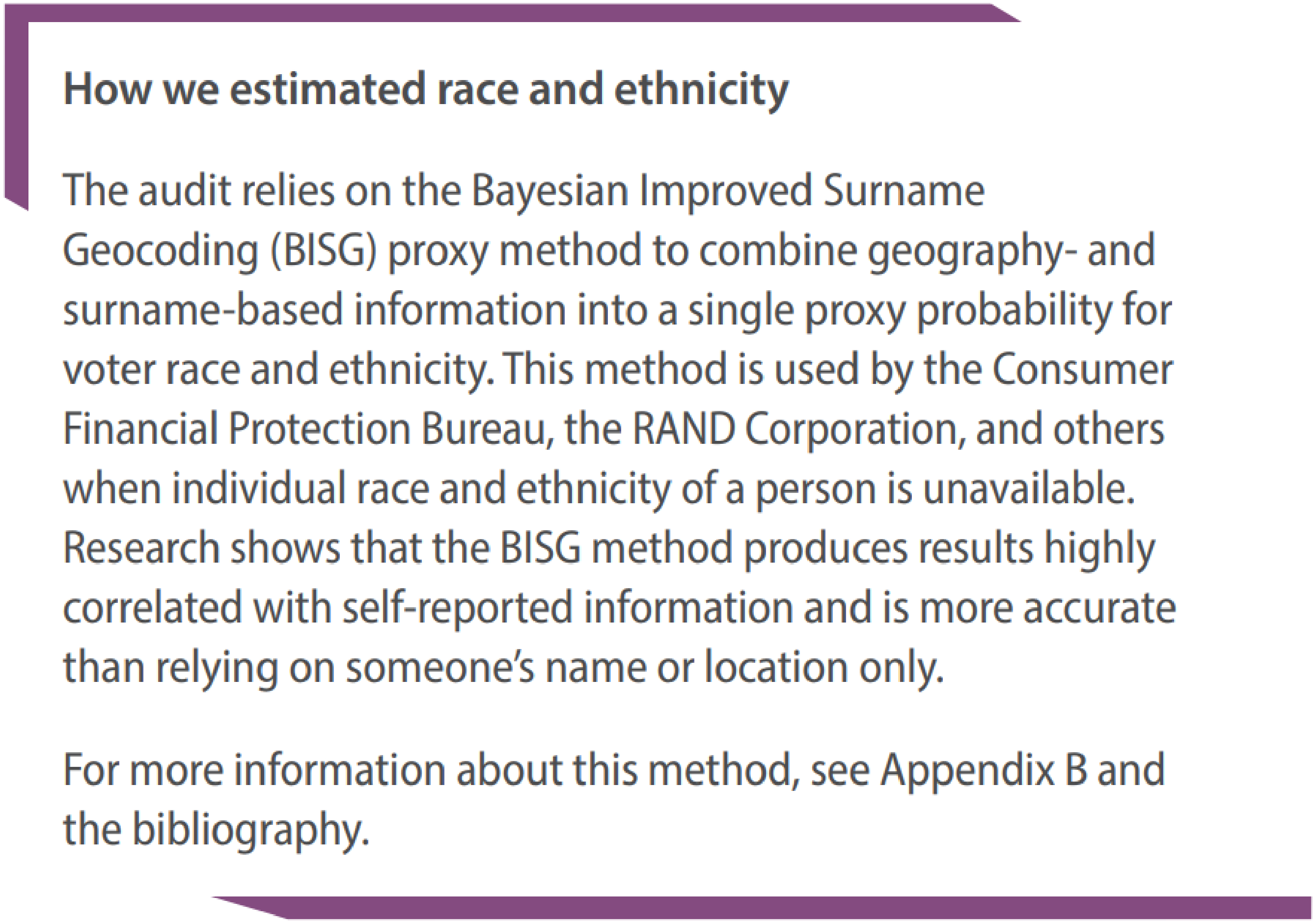
Widespread use of BISG



PERFORMANC
AUDIT

Evaluating Wasl
Ballot Rejection

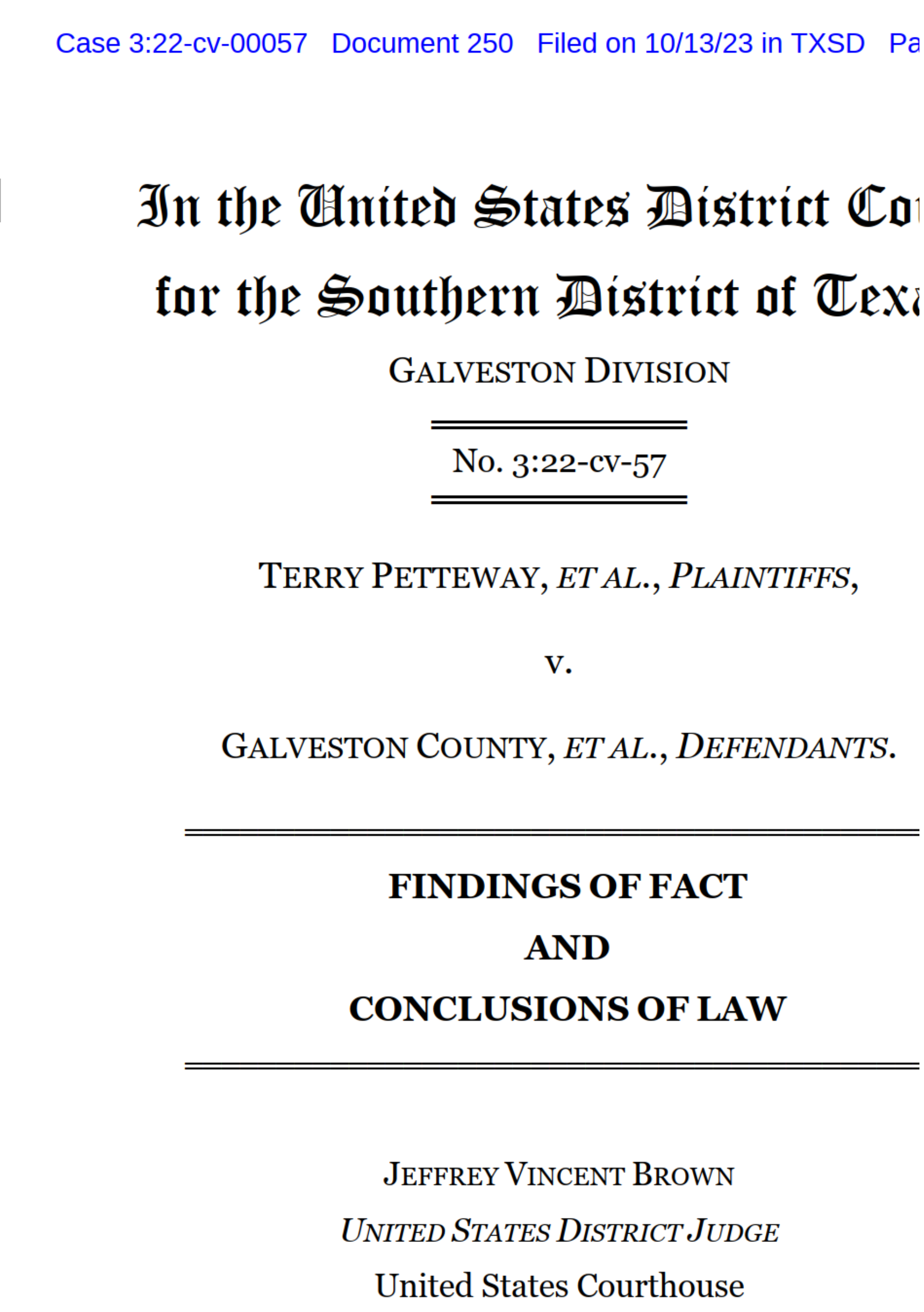
February 1, 2022



How we estimated race and ethnicity

The audit relies on the Bayesian Improved Surname Geocoding (BISG) proxy method to combine geography- and surname-based information into a single proxy probability for voter race and ethnicity. This method is used by the Consumer Financial Protection Bureau, the RAND Corporation, and others when individual race and ethnicity of a person is unavailable. Research shows that the BISG method produces results highly correlated with self-reported information and is more accurate than relying on someone's name or location only.

For more information about this method, see Appendix B and the bibliography.



In the United States District Co
for the Southern District of Tex

GALVESTON DIVISION

No. 3:22-cv-57

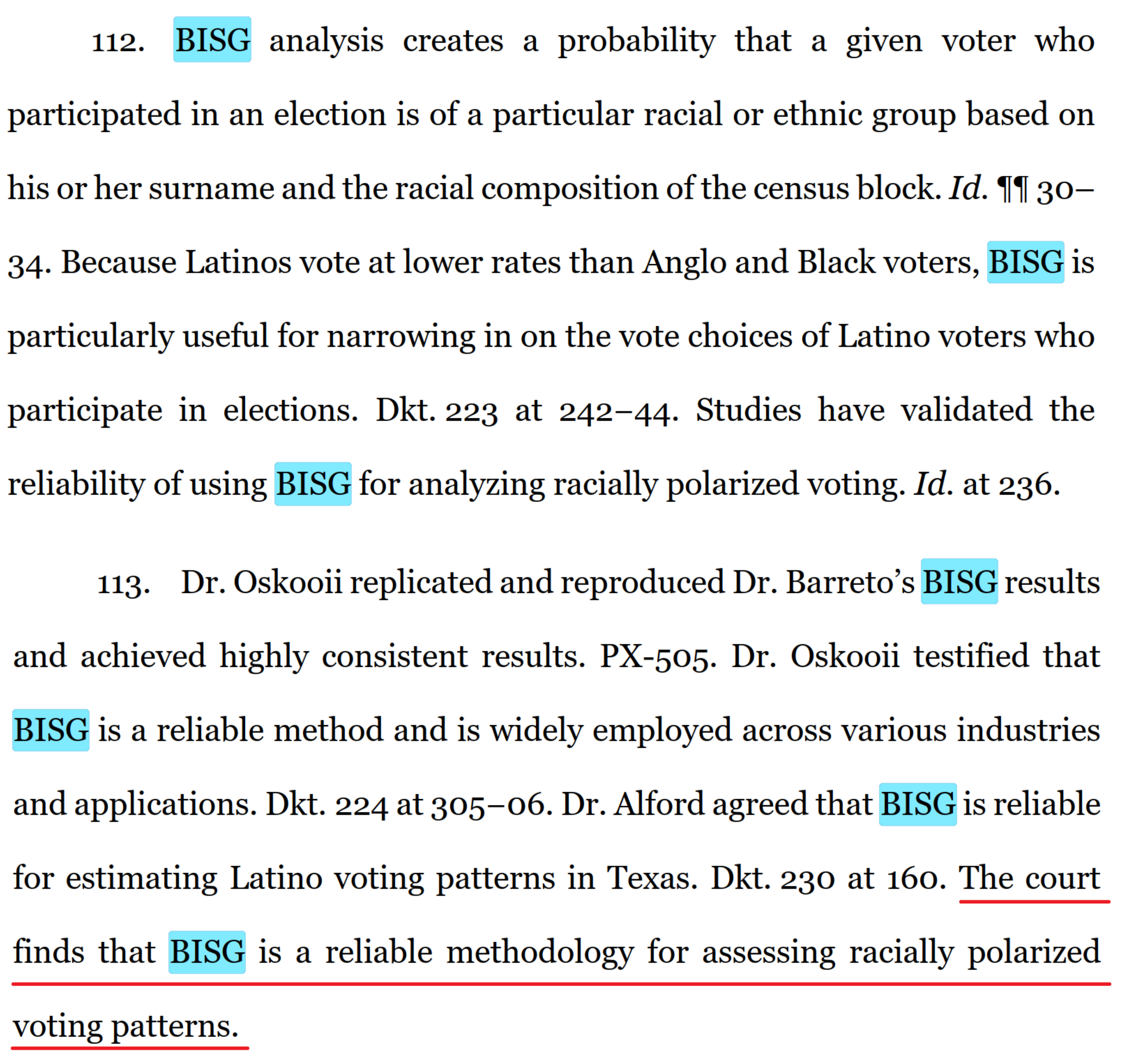
TERRY PETTEWAY, *ET AL.*, *PLAINTIFFS*,

v.

GALVESTON COUNTY, *ET AL.*, *DEFENDANTS*.

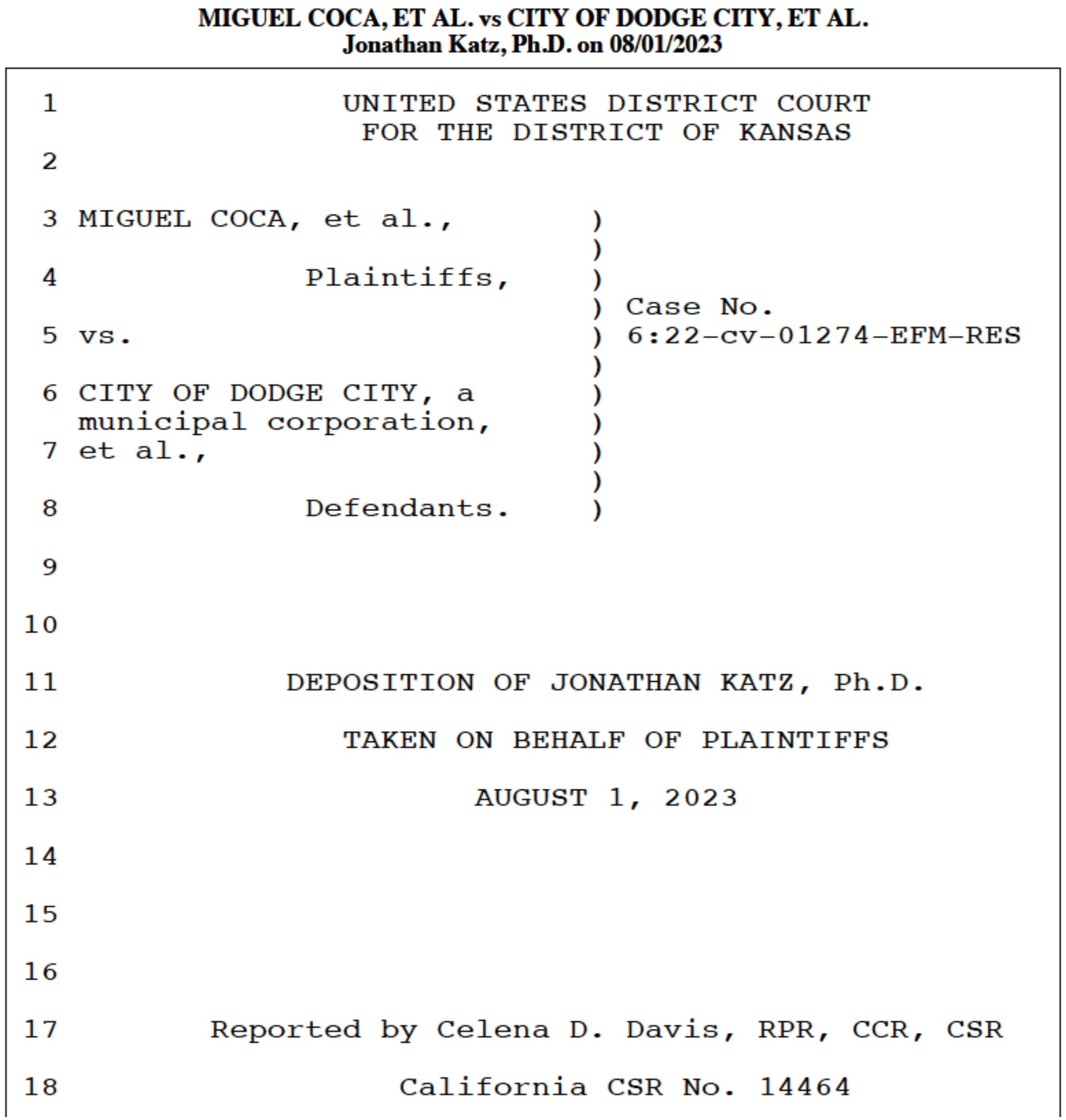
FINDINGS OF FACT
AND
CONCLUSIONS OF LAW

JEFFREY VINCENT BROWN
UNITED STATES DISTRICT JUDGE
United States Courthouse



112. BISG analysis creates a probability that a given voter who participated in an election is of a particular racial or ethnic group based on his or her surname and the racial composition of the census block. *Id.* ¶¶ 30–34. Because Latinos vote at lower rates than Anglo and Black voters, BISG is particularly useful for narrowing in on the vote choices of Latino voters who participate in elections. Dkt. 223 at 242–44. Studies have validated the reliability of using BISG for analyzing racially polarized voting. *Id.* at 236.

113. Dr. Oskooii replicated and reproduced Dr. Barreto's BISG results and achieved highly consistent results. PX-505. Dr. Oskooii testified that BISG is a reliable method and is widely employed across various industries and applications. Dkt. 224 at 305–06. Dr. Alford agreed that BISG is reliable for estimating Latino voting patterns in Texas. Dkt. 230 at 160. The court finds that BISG is a reliable methodology for assessing racially polarized voting patterns.



MIGUEL COCA, ET AL. vs CITY OF DODGE CITY, ET AL.
Jonathan Katz, Ph.D. on 08/01/2023

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF KANSAS

MIGUEL COCA, et al.,)
)
Plaintiffs,)
) Case No.
vs.) 6:22-cv-01274-EFM-RES
)
CITY OF DODGE CITY, a)
municipal corporation,)
et al.,)
)
Defendants.)

DEPOSITION OF JONATHAN KATZ, Ph.D.

TAKEN ON BEHALF OF PLAINTIFFS

AUGUST 1, 2023

Reported by Celena D. Davis, RPR, CCR, CSR

California CSR No. 14464

Q. And how common is using BISG for political scientists or for statisticians like yourself?

A. I think it's relatively common. It's -- ***when you don't have other methods or knowing respondents' race or ethnicity, it's probably one of the best things out there.***

WIDESPREAD USE OF BISG

9

Fair Lending/ Non-Discrimination

*John L. Ropiequet**

New York Attorney General's Office of Voting Rights is now using BISG to comply with new state VRA

BISG as an Advancement

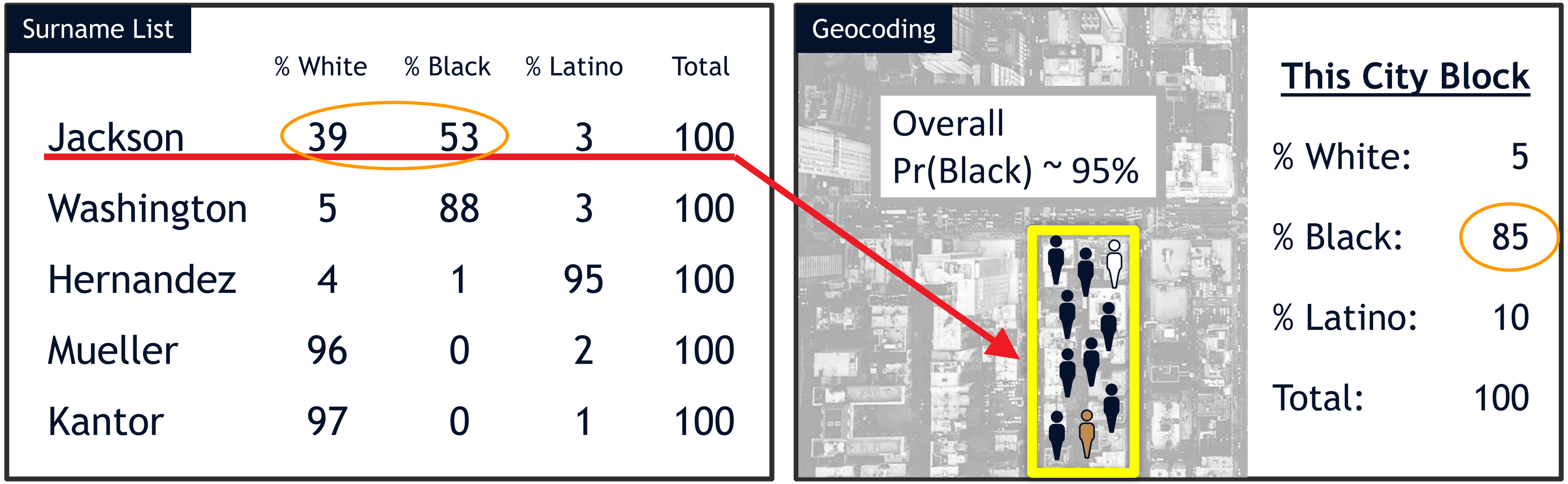
- Bayesian Improved Surname Geocoding (BISG) was developed by demographic, health, and social science experts and has been widely published and applied in the domain of public health and voting rights.

- BISG is a technique that relies on a combination of **1) census surname analysis** and **2) census block racial demographics** to provide an overall probability assessment of a voter’s race and ethnicity.

- Surname analysis has been regularly used on the voter file; however, it typically best identifies Latino and Asian American and Pacific Islander (AAPI) voters.

Example of Applied BISG

- The U.S. Census Bureau has scored all surnames against self-reported race/ethnicity and assigned probabilities to each surname in the U.S.



Surname List

	% White	% Black	% Latino	Total
Jackson	39	53	3	100
Washington	5	88	3	100
Hernandez	4	1	95	100
Mueller	96	0	2	100
Kantor	97	0	1	100

Geocoding

Overall
Pr(Black) ~ 95%

This City Block

% White:	5
% Black:	85
% Latino:	10
Total:	100

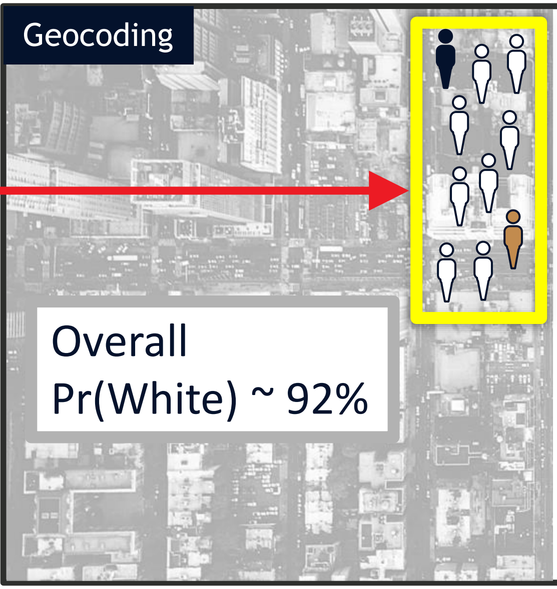
EXAMPLE OF APPLIED BISG

- BISG assigns a probability based on both surname and the census block where a voter resides

Surname List

	% White	% Black	% Latino	Total
Jackson	39	53	3	100
Washington	5	88	3	100
Hernandez	4	1	95	100
Mueller	96	0	2	100
Kantor	97	0	1	100

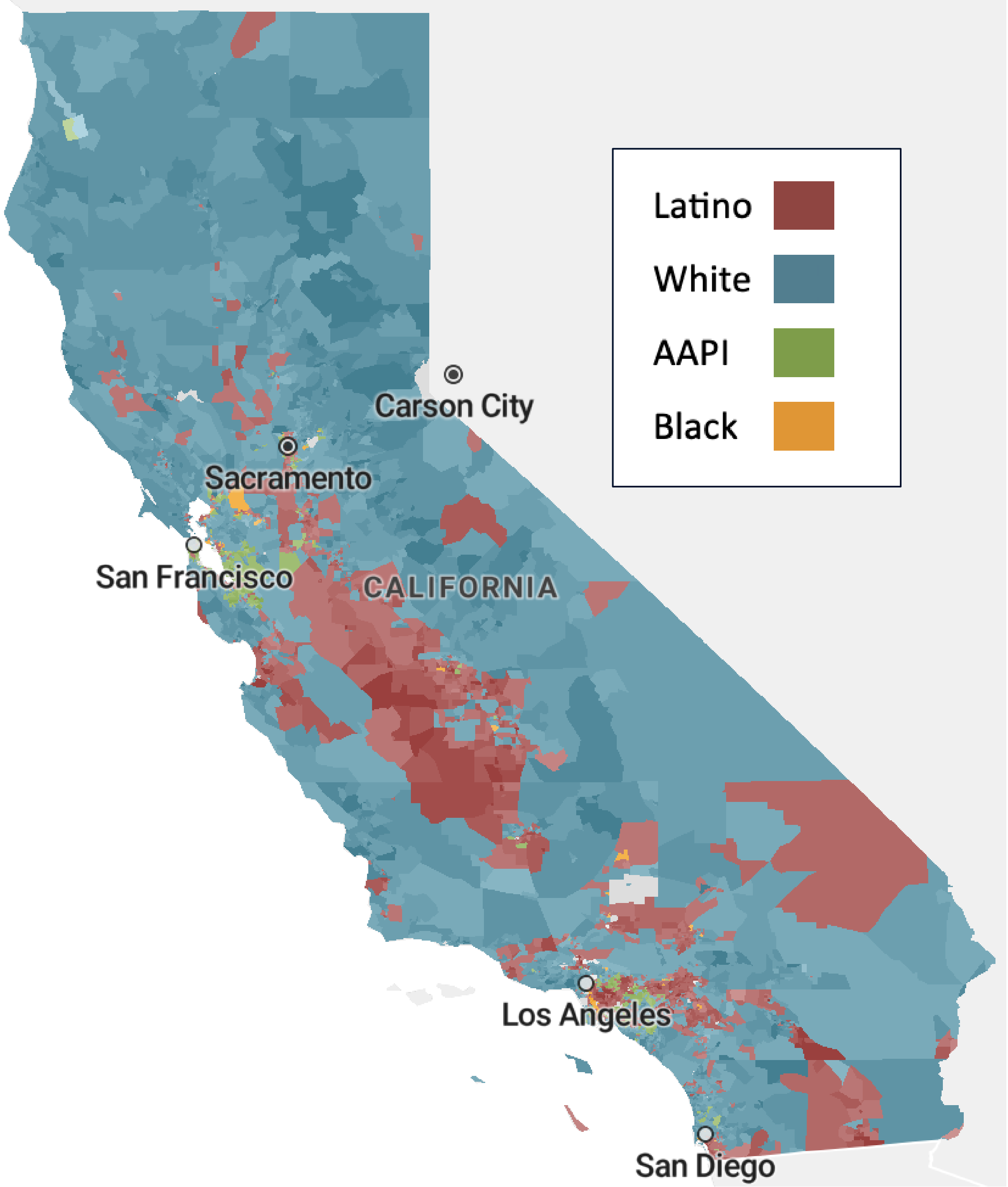
Geocoding

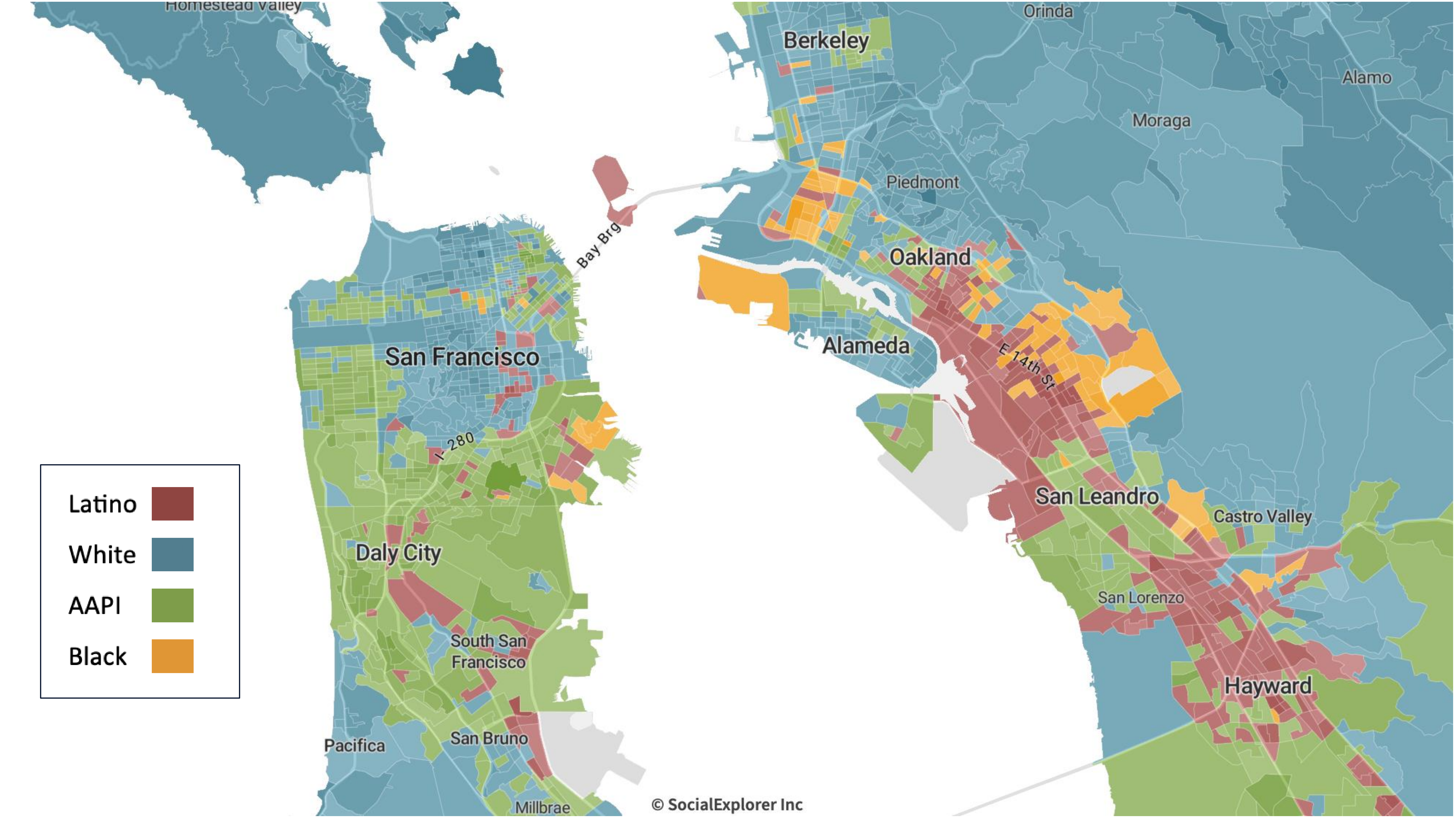


Overall
Pr(White) ~ 92%

This City Block

% White:	90
% Black:	5
% Latino:	5
Total:	100


Carson City
Sacramento
San Francisco
CALIFORNIA
Los Angeles
San Diego
Latino 🟥
White 🟦
AAPI 🟩
Black 🟧



© SocialExplorer Inc

VOTING RIGHTS PROJECT
UCLA

Research Methodology and Project Objectives

Methodology

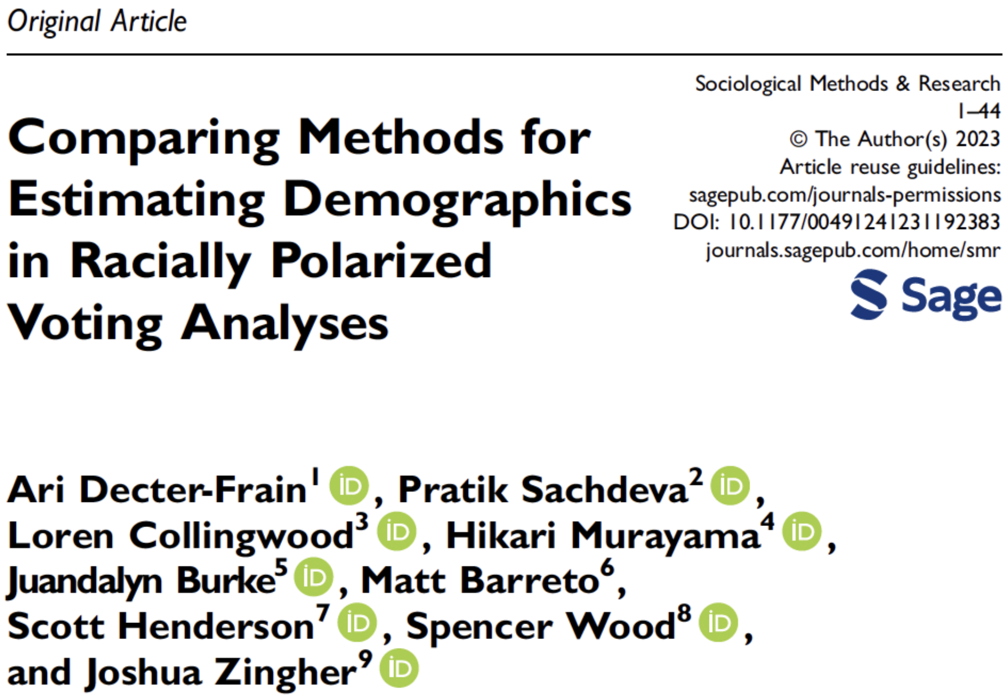
- Utilize 2020 demographic data at the census block level for the most recent and accurate data on racial/ethnic demographics within California counties.
- Include surname and first name in BISG models to improve accuracy.
- New advancements to expand BISG to use both surname and census tract data on AAPI ethnic origin groups to tabulate voting patterns for subethnic groups within the AAPI community.
- Aggregate data to geographic units such as county, city, census tract, and voting precincts.
- Validate our BISG estimates with self-reported race data on the county voter file to assess how accurate our model is performing.

Objectives

- Report important patterns such as voter registration, turnout rates, and methods of voting (mail vs. in-person) by race and ethnicity.
- Examine cross-sectional voter characteristics like age, nativity, and primary language by race and ethnicity.
- Understand voting patterns at sub-county neighborhood levels by race and ethnicity.
- Additional trends we can analyze are ballot acceptance/rejection rates and proximity to a polling location.

Note on BISG Validation

- We validate our BISG-predicted race and ethnicity estimates by comparing them to voters’ self-reported race and ethnicity.
  - Approximately 1.36 million of 5.61 million registered voters self-report their race and ethnicity.
- Using a variety of tests, we can evaluate the association between BISG-predicted and self-reported estimates.
- We find that our BISG estimates are accurate at various ecological units for the subset of voters with self-reported race and ethnicity.

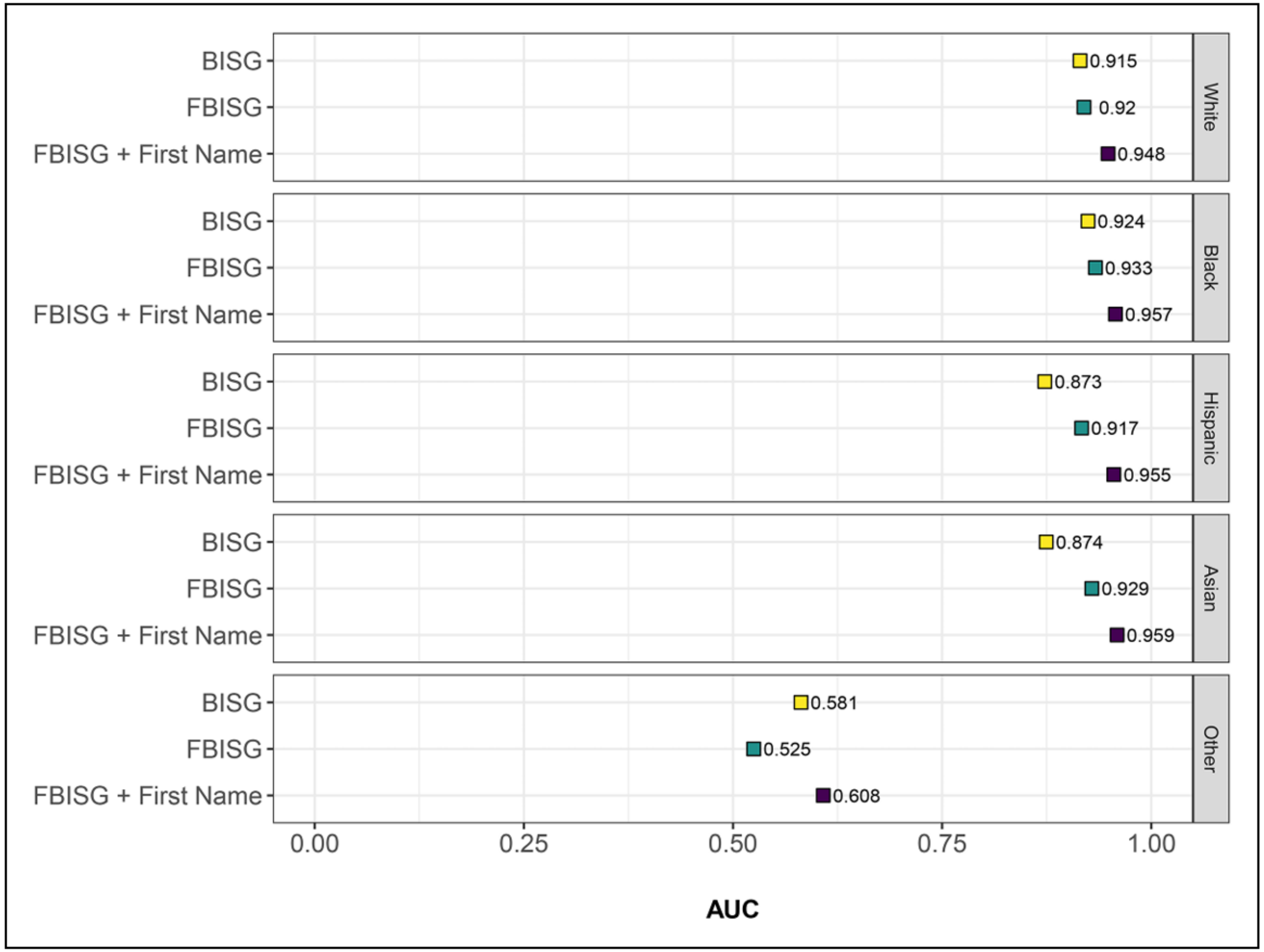


Original Article

Sociological Methods & Research
1–44
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241231192383
journals.sagepub.com/home/smr
Sage

Comparing Methods for Estimating Demographics in Racially Polarized Voting Analyses

Ari Decter-Frain[1], Pratik Sachdeva[2], Loren Collingwood[3], Hikari Murayama[4], Juandalyn Burke[5], Matt Barreto[6], Scott Henderson[7], Spencer Wood[8], and Joshua Zingher[9]



Validating BISG

BISG Compared to Self-Reporting

- We validate our BISG-predicted race and ethnicity estimates by comparing them to voters’ self-reported race and ethnicity.
  - Approximately 6 million of 22 million registered voters self-report their race and ethnicity.
- Using a variety of tests, we can evaluate the association between BISG-predicted and self-reported estimates.
- We find that our BISG estimates are accurate at various ecological units for the subset of voters with self-reported race and ethnicity.

BISG Compared to Self-Reporting on the 2022 Primary Statewide Voter File

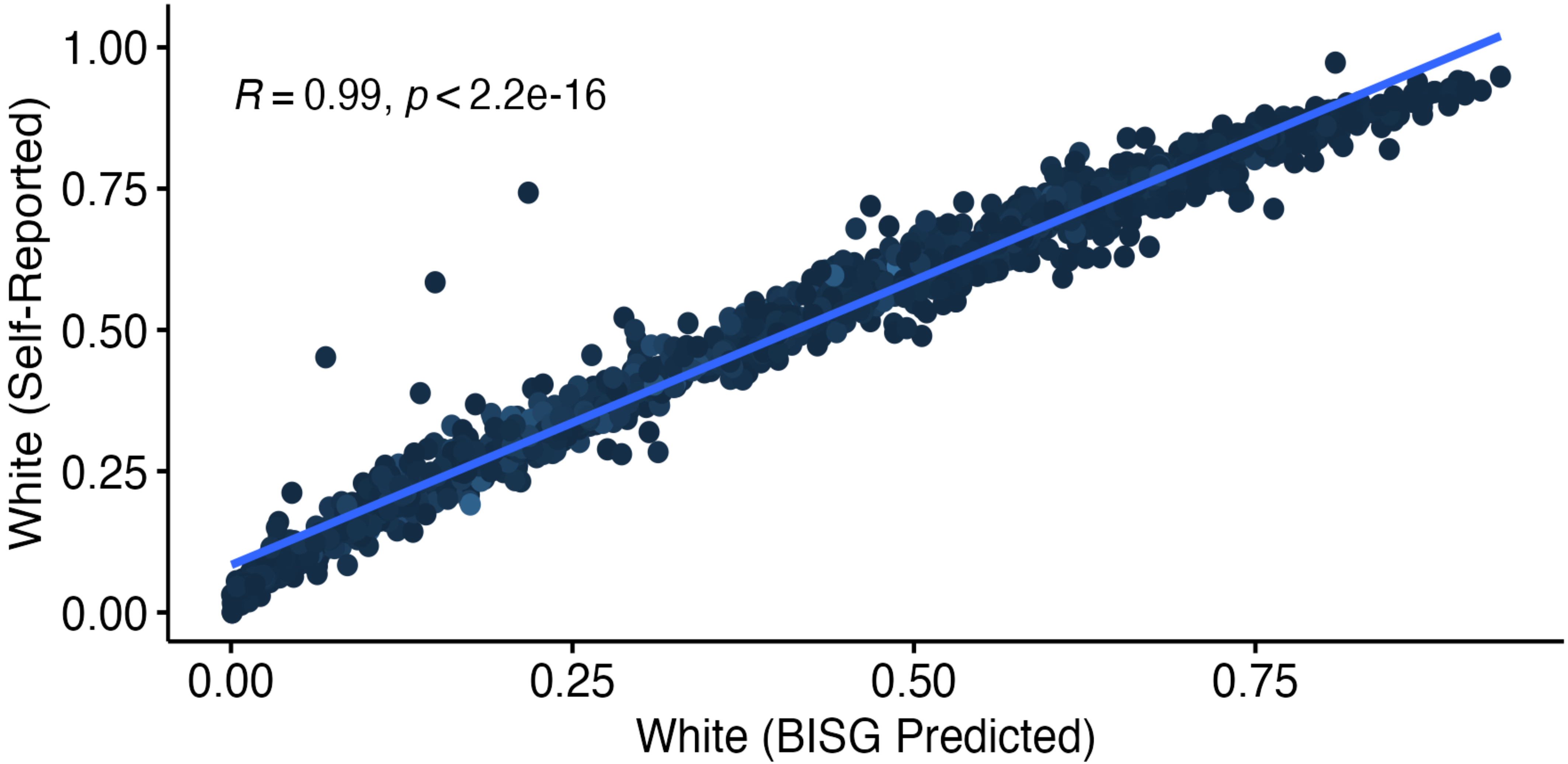
	Precinct Correlation
White	97.7%
Black	94.5%
Hispanic	97.6%
AAPI	96.7%

N = 48,022 precincts

- We subsetted the California voter file to the 6 million self-reported observations and aggregated voters into their precincts.
- Pearson’s correlation tests between a precinct’s self-reported race and ethnicity and BISG estimates show a high positive correlation for each group

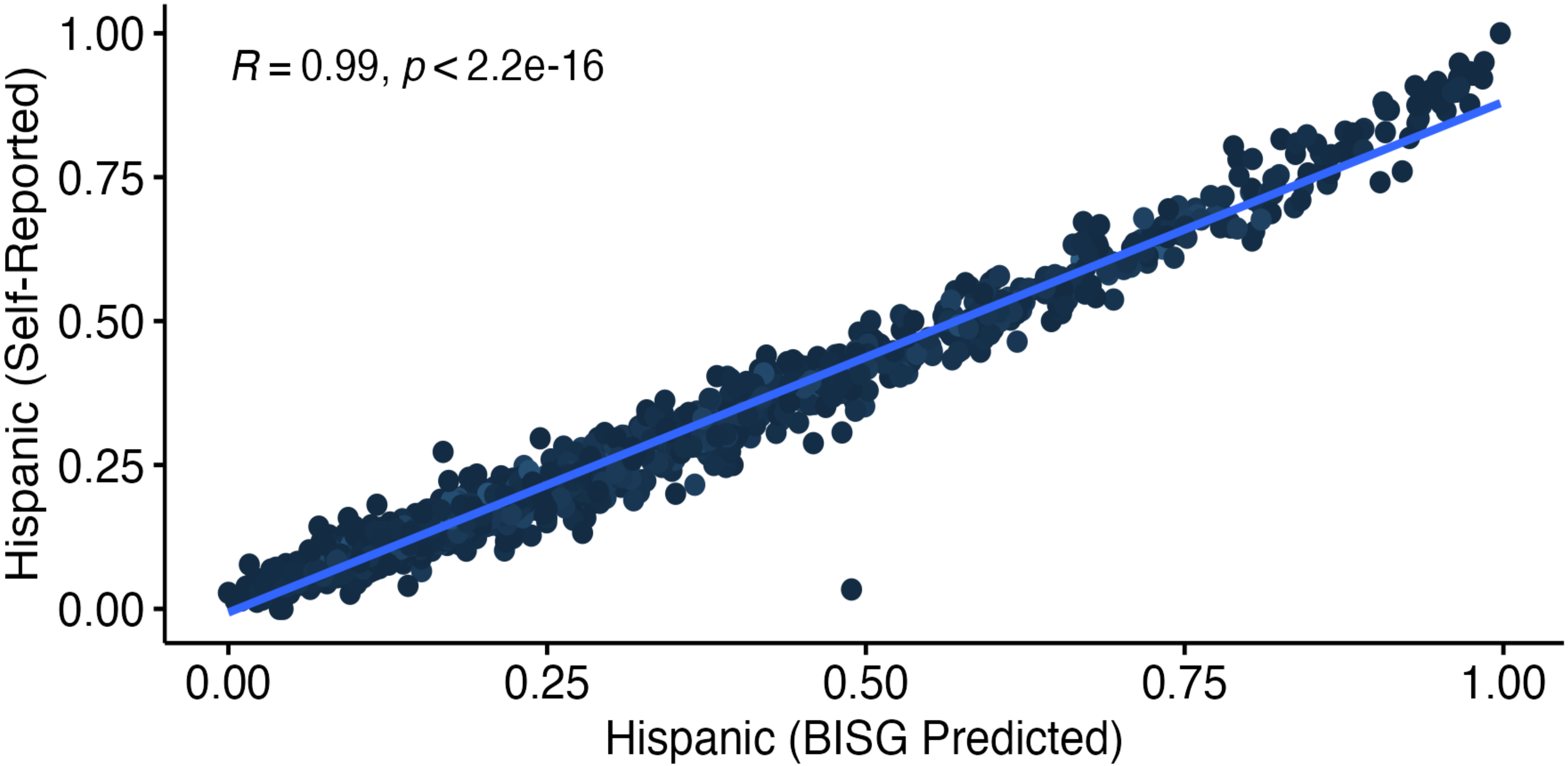
Source: California voter file extracted on January 3, 2023

BISG Prediction Compared to Self-Reporting
Orange County



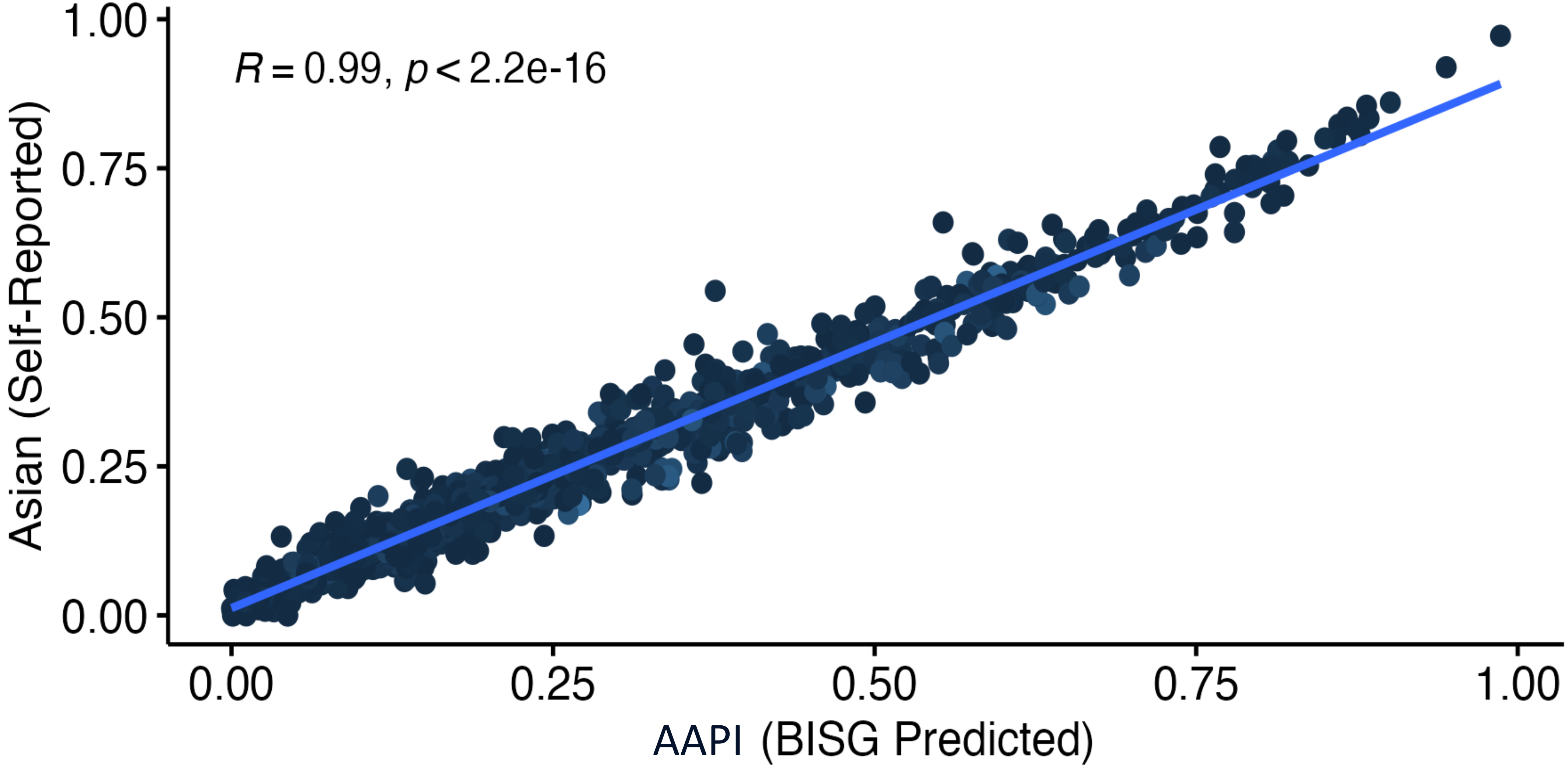
n= 1,433 precincts

BISG Prediction Compared to Self-Reporting
Orange County



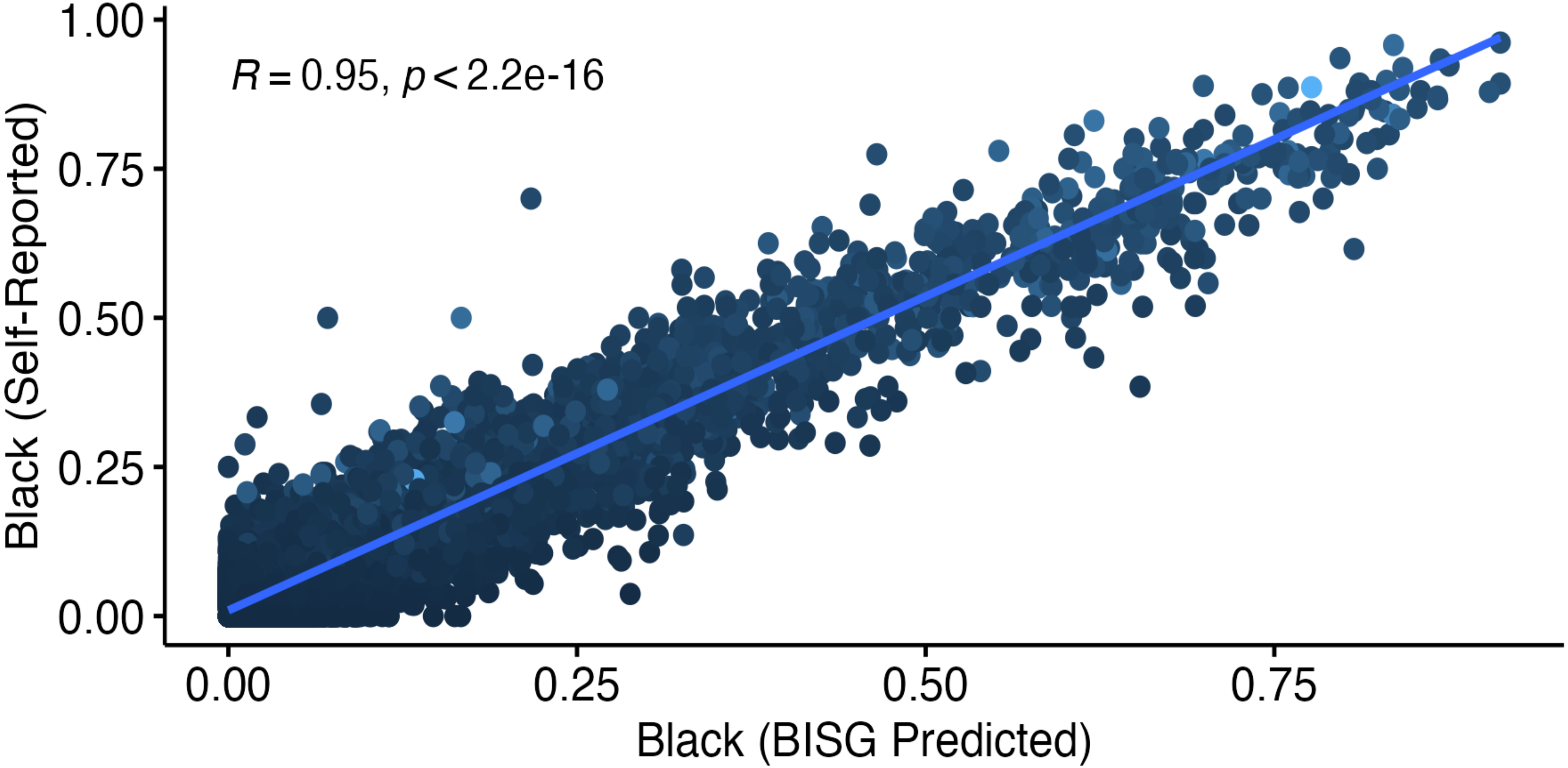
n= 1,433 precincts

BISG Prediction Compared to Self-Reporting
Orange County



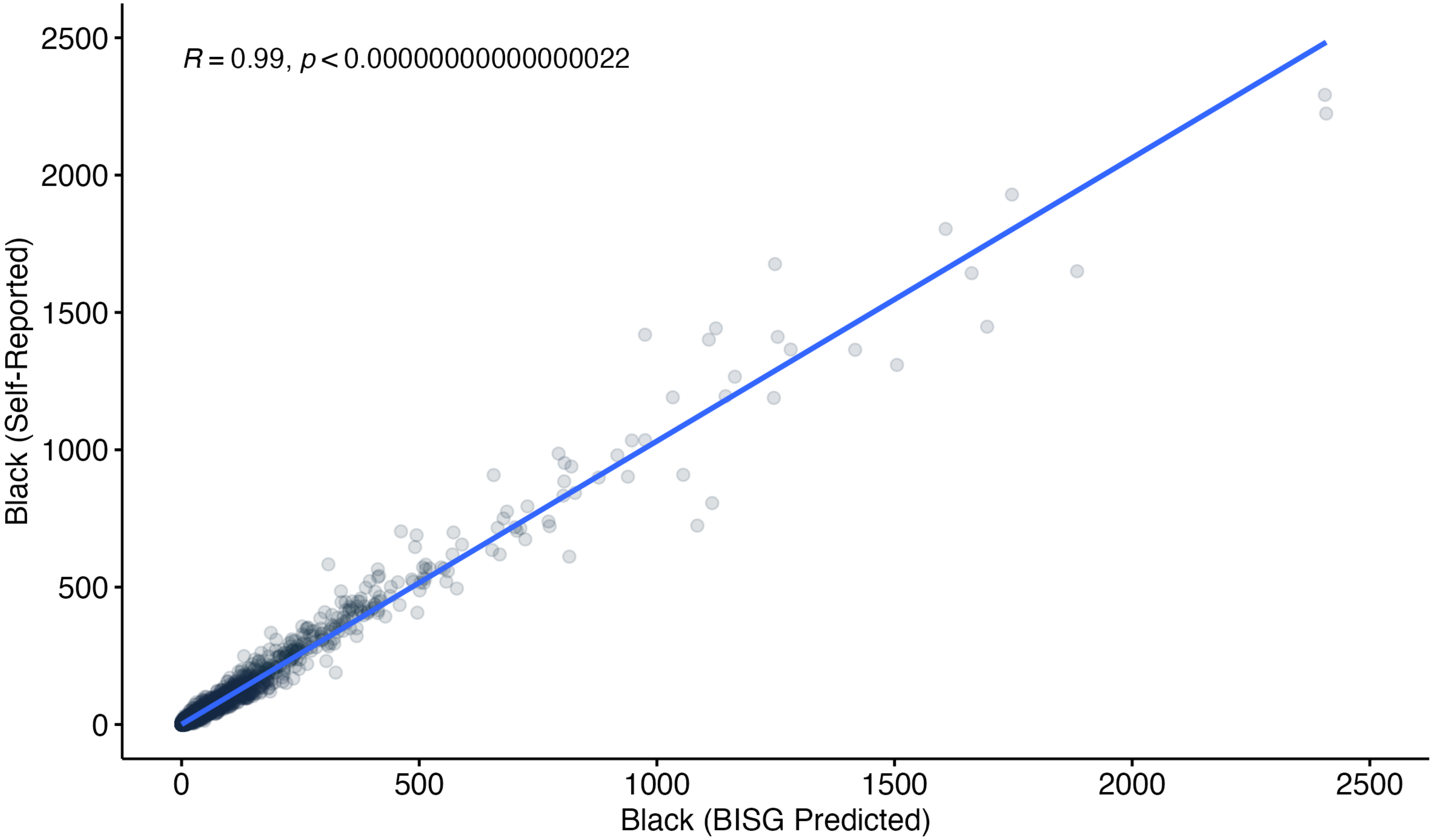
n= 1,433 precincts

BISG Prediction Compared to Self-Reporting
Los Angeles County



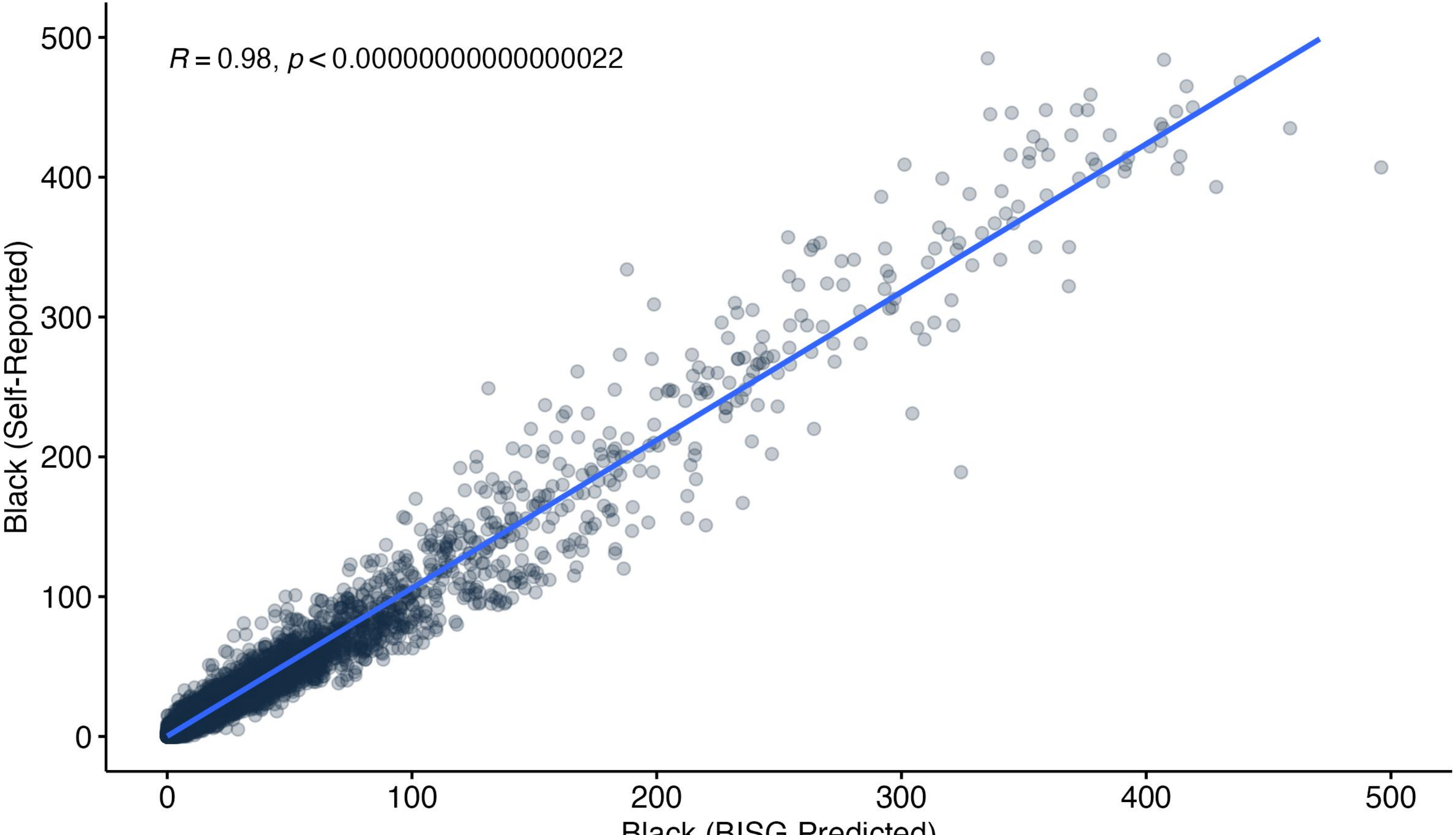
n= 22,602 precincts

BISG Prediction Compared to Self-Reporting
All California Counties



BISG Prediction Compared to Self-Reporting

(Precincts w/ Over 1,000 Black Voters)



VOTING RIGHTS PROJECT UCLA

Utilizing Bayesian Improved Surname Geocoding (BISG)

Registered Voters by Race and Ethnicity in the 2022 Elections

		Total	White		Latino		AAPI		Black		Other	
Primary	Statewide	21,902,922	8,703,339	39.7%	6,995,224	31.9%	3,690,502	16.8%	1,151,871	5.3%	1,361,986	6.2%
	VCA	16,742,926	6,478,422	38.7%	5,359,839	32.0%	3,038,629	18.1%	877,007	5.2%	989,028	5.9%
	Non-VCA	5,159,996	2,224,917	43.1%	1,635,385	31.7%	651,872	12.6%	274,864	5.3%	372,958	7.2%
General	Statewide	21,952,201	8,665,311	39.5%	7,057,357	32.1%	3,730,176	17.0%	1,147,219	5.2%	1,352,138	6.2%
	VCA	16,774,973	6,448,846	38.4%	5,402,113	32.2%	3,068,170	18.3%	873,562	5.2%	982,282	5.9%
	Non-VCA	5,177,228	2,216,465	42.8%	1,655,244	32.0%	662,006	12.8%	273,657	5.3%	369,855	7.1%

Turnout Rate by Race and Ethnicity in the 2022 Elections

		Total	White	Latino	AAPI	Black	Other
Primary	Statewide	33.0%	44.4%	21.2%	29.9%	28.0%	33.3%
	VCA	33.1%	44.7%	21.6%	29.6%	28.8%	33.3%
	Non-VCA	32.9%	43.8%	19.7%	31.4%	25.7%	33.3%
General	Statewide	50.4%	64.3%	36.0%	47.3%	43.5%	51.4%
	VCA	50.3%	64.4%	36.2%	47.0%	43.7%	51.2%
	Non-VCA	50.8%	63.8%	35.5%	48.6%	42.7%	52.0%

- Turnout for Latino and Black voters in the primary and general elections is higher in VCA counties than in non-VCA counties.

Participation Method by Race and Ethnicity in the 2022 Elections

		Total	White	Black	Latino	APPI	Other
Primary	Vote-By-Mail	91.6%	92.1%	89.8%	89.2%	93.6%	91.3%
	Vote Center	6.5%	5.9%	8.2%	8.8%	5.2%	6.4%
	Polling Place	1.8%	1.9%	1.8%	1.8%	1.1%	2.2%
General	Vote-By-Mail	88.1%	89.0%	86.4%	85.2%	90.3%	87.7%
	Vote Center	9.6%	8.6%	11.2%	12.6%	8.0%	9.4%
	Polling Place	1.9%	2.1%	1.9%	1.7%	1.2%	2.5%

Registered Voters by AAPI Subethnic Group in the 2022 Elections

	Total AAPI	Chinese	Filipino	Indian	Vietnamese	Korean	Japanese	Other
Primary	3,690,502	970,122	611,171	539,170	525,588	293,356	163,069	588,025
General	3,730,175	942,487	647,687	527,743	506,150	290,447	143,990	671,672

Future Objectives

- Continue to improve our BISG models by incorporating additional demographic data.
  - Test citizen voting-age population (CVAP) data at the block group level to improve predicted Latino estimates.
- Build a database of polling locations and ballot dropboxes to further evaluate equity in the distance to cast a ballot.

VOTING RIGHTS PROJECT UCLA

latino.ucla.edu/votingrights/