

Estimating Racial and Ethnic Disparities When Group Membership Is Unavailable

Kosuke Imai

Harvard University

Seminar at Department of Political Science

Washington University in St. Louis

October 15, 2026

Joint work with Kyla Chasalow, Noah Dasanaïke, and Cory McCartan

Motivation

- Importance of racial disparity estimation in many fields:
public health, employment, voting, criminal justice, taxation, housing, lending, etc.
- But, often individual race information is not available
 - law may prohibit collection of information about race (e.g., Equal Credit Opportunity Act)
 - agencies and companies may not wish to collect such information
- How should we estimate racial disparities when race is not observed?
- Standard approach:
 - 1 Bayesian Improved Surname Geocoding (**BISG**) for predicting race
 - 2 Downstream analysis using predicted race (classification or weighting)
- Today's talk
 - 1 How to conduct downstream analysis without bias (**BIRDIE**)
 - 2 How to estimate calibrated racial probabilities from group-specific name lists (**ℓBISG**)

2020 United States Census Data

- Name-race frequency data:
 - all first (53,615) and last (156,621) names that occur 100 or more times
 - 6 racial categories: White, Black, Hispanic, Asian, AIAN, Multi-racial
- Race-geography data:
 - counts of 6 racial categories by census geography
 - counties, tracts, blocks, etc.
- Name-race frequency data are typically not available
 - ① outside of the US
 - ② for granular racial and ethnic groups even in the US
- Race-geography data are typically available but may be lacking in some cases
- Possible elimination of racial questions in the 2030 Census

BISG (Bayesian Improved Surname Geocoding; Elliott et al. (2009))

- How to predict race using an individual's surname and address?
- Probabilistic prediction: $\mathbb{P}(R | S, G)$ with Race, Surname, and Geolocation
- Census data: $\mathbb{P}(R, S)$, $\mathbb{P}(R, G)$
- Assumption: surname is independent of geolocation given race

$$S \perp\!\!\!\perp G \mid R$$

- If R represents a finer category, the assumption may be violated
- Bayes rule:

$$\mathbb{P}(R = r \mid S = s, G = g) = \frac{\overbrace{\mathbb{P}(S = s \mid R = r)}^{\text{surname prevalence}} \times \overbrace{\mathbb{P}(R = r \mid G = g)}^{\text{racial prevalence}}}{\underbrace{\sum_{r'} \mathbb{P}(S = s \mid R = r') \times \mathbb{P}(R = r' \mid G = g)}_{\text{normalization so that probabilities add up to 1}}}$$

- BIFSG: we can incorporate first name data in a similar way

How to conduct downstream analysis without bias (McCartan et al. 2025. *JASA*)

Good Race Prediction Can Lead to Biased Racial Disparity Estimates

- Estimate racial disparities $\mu_{Y|R}(y | r) = \mathbb{P}(Y = y | R = r)$
- **weighting**: use the BISG probabilities as weights

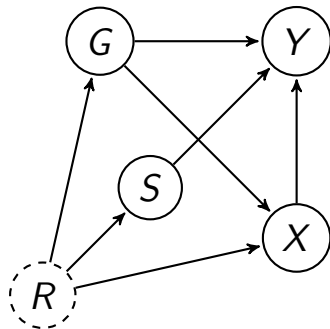
$$\hat{\mu}_{Y|R}^{\text{wtd}}(y | r) = \frac{\sum_i \mathbf{1}\{Y_i = y\} \hat{P}_{ir}}{\sum_i \hat{P}_{ir}}$$

- Bias of the weighted estimator (Chen *et al.* 2019)

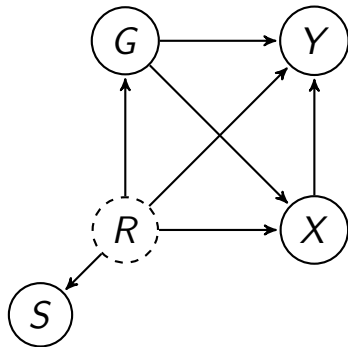
$$-\frac{\mathbb{E}[\text{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = r\} | G, X, S)]}{\mathbb{P}(R_i = r)}$$

- Likely to lead to an underestimation of disparity
- Required assumption:

$$Y \perp\!\!\!\perp R | G, S, X$$



More Credible Identification Strategy



- Alternative assumption:

$$Y \perp\!\!\!\perp S \mid G, R, X$$

- Surname as a proxy for race
- Race can directly or indirectly affects the outcome
- Potential violations:
 - name-based discrimination
 - coarse racial categories
- Anonymous application

Surname as a High-dimensional Proxy for Race

- Identification (Kuroki and Pearl, 2014)

$$\begin{aligned} & \overbrace{\mathbb{P}(Y_i = y \mid G_i = g, X_i = x, S_i = s)}^{\text{observed data}} \\ &= \sum_{r \in \mathcal{R}} \underbrace{\mathbb{P}(Y_i = y \mid R_i = r, G_i = g, X_i = x)}_{\text{unknown parameters}} \times \underbrace{\mathbb{P}(R_i = r \mid G_i = g, X_i = x, S_i = s)}_{\text{BISG probability}} \end{aligned}$$

- $(|\mathcal{Y}| - 1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{S}|$ equations
- $(|\mathcal{Y}| - 1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{R}|$ unknown parameters
- OLS estimator (see also Fong and Tyler, 2021):

$$\hat{\mu}_{Y|RGX}^{(\text{ols})}(y \mid \cdot, g, x) = (\hat{\mathbf{P}}_{\mathcal{I}(xg)}^\top \hat{\mathbf{P}}_{\mathcal{I}(xg)})^{-1} \hat{\mathbf{P}}_{\mathcal{I}(xg)} \mathbb{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\},$$

- compute this for each g and x , and aggregate
- unbiased estimate of $\mathbb{P}(Y_i = y \mid R_i = r)$
- ignores the fact that $\mathbb{P}(Y_i = y \mid R_i = r, G_i = g, X_i = x)$ is probability
- inefficient if $|\mathcal{G}| \times |\mathcal{X}|$ is large

BIRDiE (Bayesian Instrumental Regression for Disparity Estimation)

- Flexible and scalable probabilistic model that integrates BISG
- Posterior:

$$\pi(\Theta, \mathbf{R} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \propto \pi(\Theta) \prod_{i=1}^N \underbrace{\pi(Y_i \mid R_i, G_i, X_i, \Theta)}_{\text{complete-data model}} \underbrace{\pi(R_i \mid G_i, X_i, S_i)}_{\text{BISG prob. } \hat{P}_{ir}}$$

- Models:

- 1 Complete-pooling:

$$Y_i \mid R_i, G_i, X_i, \Theta \sim \text{Cat}_Y(\boldsymbol{\theta}_{R_i}), \quad \boldsymbol{\theta}_r \stackrel{iid}{\sim} \text{Dir}(\boldsymbol{\alpha})$$

- 2 Saturated (no pooling):

$$Y_i \mid R_i, G_i, X_i, \Theta \sim \text{Cat}_Y(\boldsymbol{\theta}_{R_i G_i X_i}), \quad \boldsymbol{\theta}_{rgx} \stackrel{iid}{\sim} \text{Dir}(\boldsymbol{\alpha})$$

- 3 Partial pooling (mixed effects): $\mathbf{W}$ group-level covariates, $\mathbf{Z} = (X, G)$

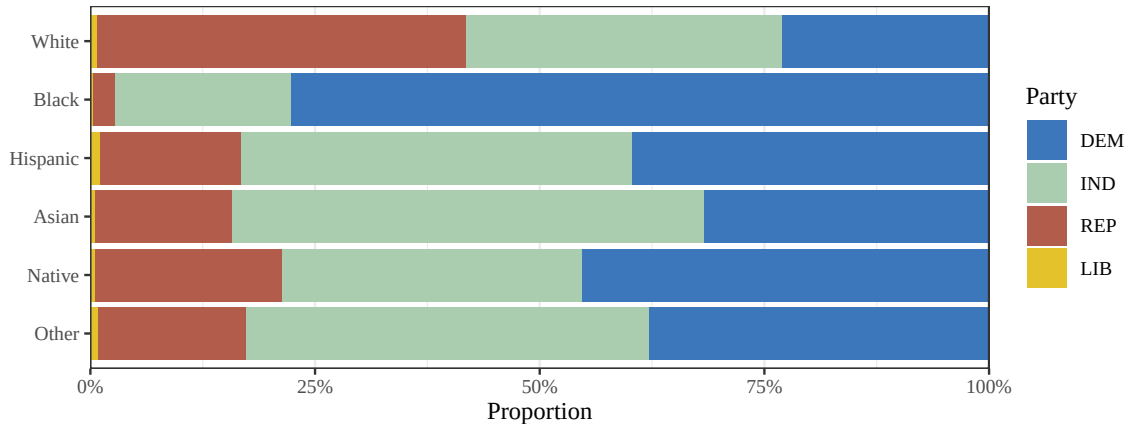
$$Y_i \mid R_i, G_i, X_i, \Theta \sim \text{Cat}_Y(g^{-1}(\boldsymbol{\mu}_{rgx})), \quad \boldsymbol{\mu}_{rgxy} = \mathbf{W}\boldsymbol{\beta}_{ry} + \mathbf{Z}\mathbf{u}_{ry}$$

$$\mathbf{u}_{ry} \mid \phi_{ry} \sim \mathcal{N}(0, \Sigma(\phi_{ry})), \quad \boldsymbol{\beta}_{ry} \stackrel{iid}{\sim} f_{\beta}, \quad \phi_{ry} \stackrel{iid}{\sim} f_{\phi}$$

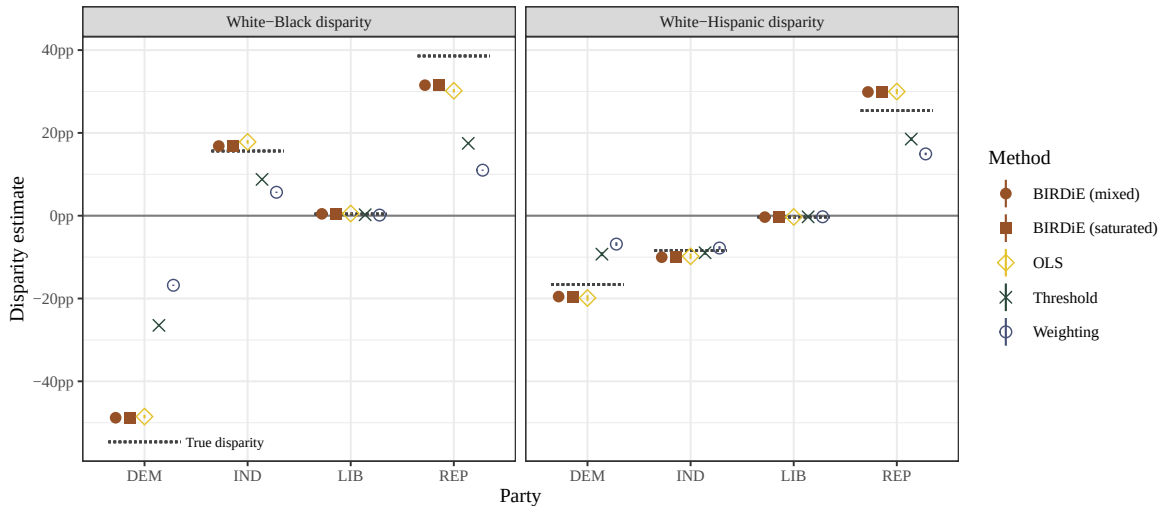
- EM algorithm / Gibbs sampler: improved race probabilities

Empirical Validation

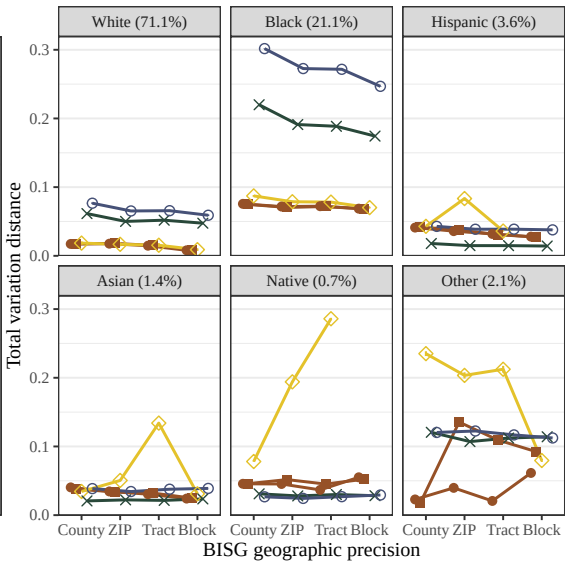
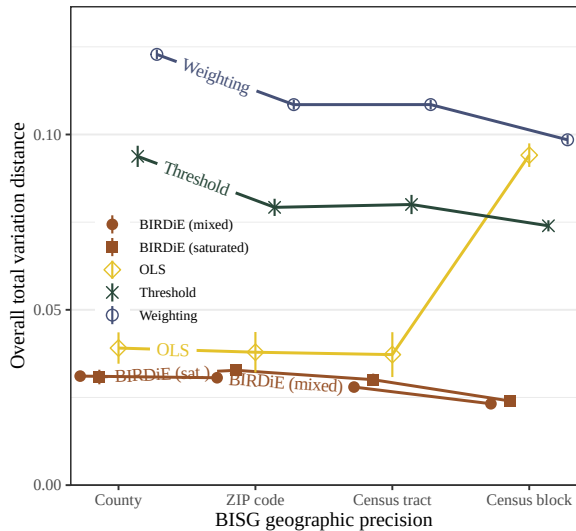
- 2022 North Carolina voter file: 5.8 million voters with self-reported race
- Subset 1 million voters $\rightsquigarrow$ negligible sampling uncertainty
- Focus on party registration



Estimates of Racial Disparity in Party Registration



Total Variation Distance



How to estimate calibrated probabilities from group-specific name lists

(Chasalow *et al.* arXiv preprint)

Fundamental Limitation of BISG Methodology

- BISG is only applicable when name-race frequency data are available
 - such data are not available for uncommon names and smaller ethnic categories
 - difficult to obtain them outside of the US
- Can we use name lists instead?
 - name lists are available in many contexts
 - LLM can quickly generate them
 - but, unknown bias due to misclassification remains
- Can we make valid probabilistic race-ethnicity predictions from group-specific name lists?

embedding-powered BISG (eBISG; Daisanaike and Imai (2026))

- Name embeddings = vector-valued numeric representations of names
 - No new assumption: $E(S) \perp\!\!\!\perp G \mid R$
 - We use open-source Multilingual E5-Large model from Microsoft
 - Other models give similar results for our applications
- We can use machine learning with name embeddings
 - Train neural network using the name-race frequency data with common names
 - Predict race probabilities for uncommon names
- eBISG:

$$\mathbb{P}(R = r \mid E = e, G = g) \propto \underbrace{\mathbb{P}(E = e \mid R = r)}_{= \frac{\mathbb{P}(R=r|E=e) \mathbb{P}(E=e)}{\mathbb{P}(R=r)}} \times \mathbb{P}(R = r \mid G = g)$$

- But, this still requires the name-race frequency data

Identification

- BISG / eBISG assumption implies

$$\mathbb{P}(E = e \mid G = g) = \sum_{r=1}^{|\mathcal{R}|} \mathbb{P}(E = e \mid R = r) \mathbb{P}(R = r \mid G = g)$$

- We can identify without an additional assumption so long as $|\mathcal{G}| \geq |\mathcal{R}|$

$$\phi_e = (\mathbf{M}^\top \mathbf{M})^{-1} \mathbf{M}^\top \rho_e$$

where $\phi_{e,r} = \mathbb{P}(E = e \mid R = r)$, $\mathbf{M}_{g,r} = \mathbb{P}(R = r \mid G = g)$, and $\rho_{e,g} = \mathbb{P}(E = e \mid G = g)$

- But, embeddings are high-dimensional and $\mathbb{P}(E = e \mid G = g)$ is generally degenerate

list-powered BISG (ℓ BISG)

- We **cluster** embeddings to estimate: $\mathbb{P}(B(E) = b \mid G = g)$
- No new assumption: $B(E) \perp\!\!\!\perp G \mid R$

$$\mathbb{P}(B(E) = b \mid G = g) = \sum_{r=1}^{|\mathcal{R}|} \mathbb{P}(B(E) = b \mid R = r) \mathbb{P}(R = r \mid G = g)$$

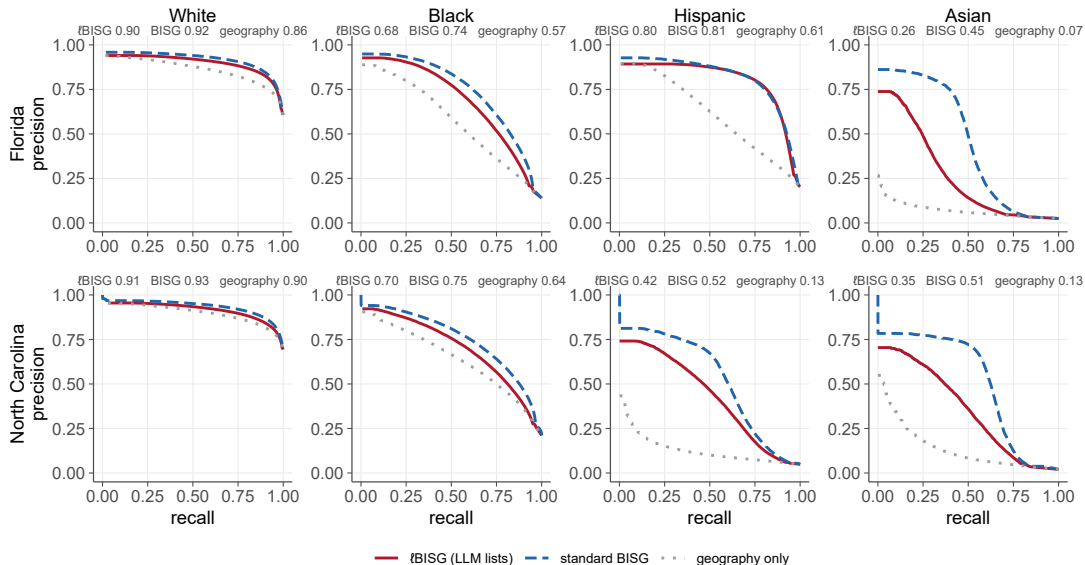
- Leverage group-specific name lists: L_r is an indicator of being on the name list for group r
- ℓ BISG algorithm
 - 1 Train a classifier to estimate $f_{r,e} = \mathbb{P}(L_r = 1 \mid E = e)$ for each r and given e
 - 2 Apply the k -means algorithm to the logit of $\hat{\mathbf{f}}_e = (\hat{f}_{1,e}, \dots, \hat{f}_{|\mathcal{R}|,e})$ into K clusters
 - 3 Run regression $\hat{\phi}_b = (\hat{\mathbf{M}}^\top \hat{\mathbf{M}})^{-1} \hat{\mathbf{M}}^\top \hat{\rho}_b$ using the identification result
- Choose K through sample-splitting to maximize

$$Q(K) = \frac{1}{|\mathcal{I}_{\text{val}}|} \sum_{i \in \mathcal{I}_{\text{val}}} \left[\log \hat{\mathbb{P}}_{\text{train}}(B = B_i \mid G = G_i) - \log \hat{\mathbb{P}}_{\text{train}}(B = B_i) \right]$$

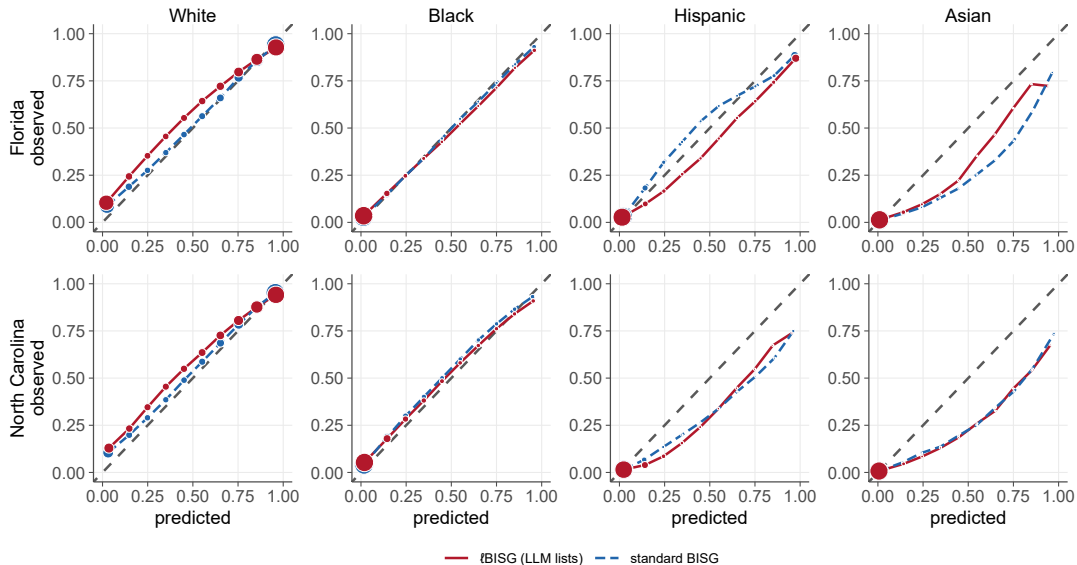
Validation I: Comparison between ℓ BISG and Standard BISG

- Florida and North Carolina voter files with self-reported race
- Use LLM to generate lists of 1000 names for Whites, Blacks, Hispanics, and Asians
 - ① prompt Gemini 3.0 Flash six times at the temperature of 1.0
 - ② stratified by national origin (e.g., English, German, French etc. for Whites)
 - ③ 300 names per call (6 calls) and take top 1,000 that appear for most calls
- Use tracts as geography, but the results are not sensitive to the length of name lists
- Outcome: Share of registered Democrats

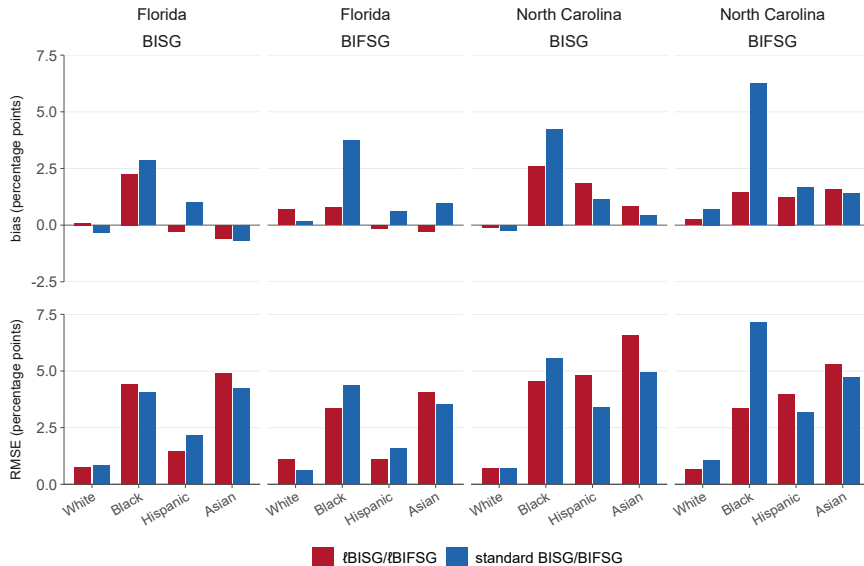
ℓBISG Prediction is as Accurate as Standard BISG



ℓ BISG Prediction is as Calibrated as Standard BISG



ℓ BISG Performs as well as Standard BISG for Downstream Analysis



ℓ BISG without Race-Geography Data

- In some cases, neither race-name frequency data nor race-geography data are unavailable
- **Assumption 1:** list construction must not depend on geography

$$L_r \perp\!\!\!\perp G \mid R$$

- The earlier identification strategy does not work since $\mathbb{P}(R \mid G)$ is unknown

$$\mathbb{P}(L_{r'} = 1 \mid G) = \sum_{r=1}^{|\mathcal{R}|} \mathbb{P}(L_{r'} \mid R = r) \mathbb{P}(R = r \mid G), \quad \text{and} \quad \sum_{r=1}^{|\mathcal{R}|} \mathbb{P}(R = r \mid G) = 1$$

- **Assumption 2:** lists are exclusive, i.e., $\mathbb{P}(L_{r'} = 1 \mid R = r) = 0$ for all $r' \neq r$
- Under these additional assumptions, we have,

$$\mathbb{P}(L_r = 1 \mid G) = \mathbb{P}(L_r \mid R = r) \mathbb{P}(R = r \mid G), \quad \text{and} \quad \sum_{r=1}^{|\mathcal{R}|} \mathbb{P}(R = r \mid G) = 1$$

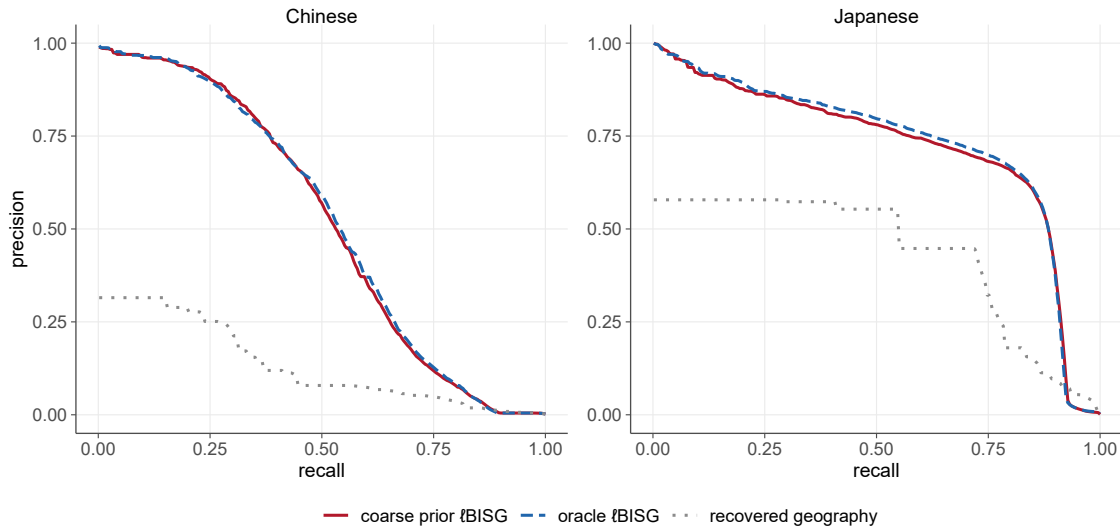
which identify $\mathbb{P}(R = r \mid G)$, making it possible to apply ℓ BISG as before

- Extension to the setting where $P(R \mid G)$ is available for a coarse group

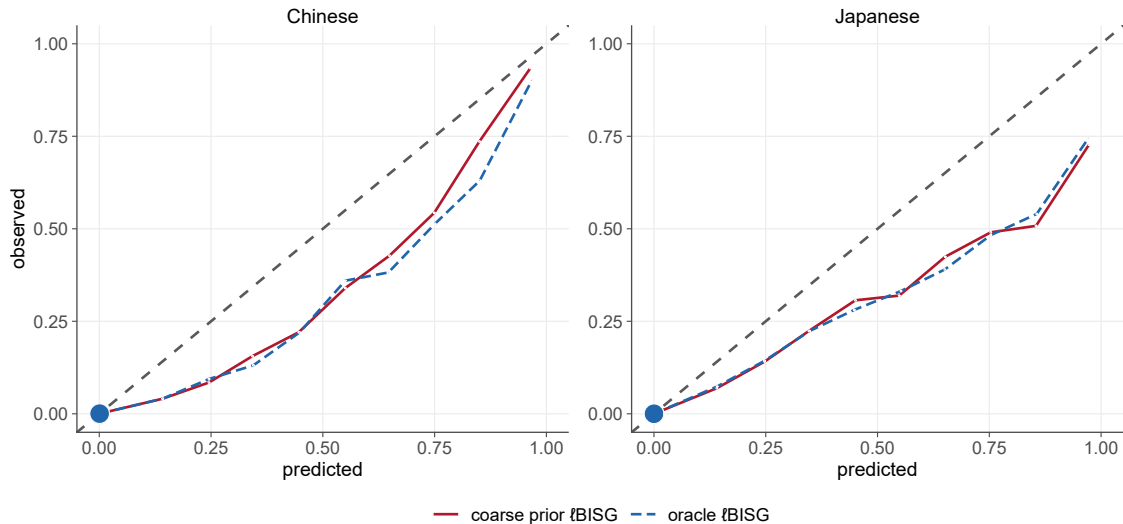
Validation II: Predicting Asian Subgroups Using Historical US Census

- 1900 Census includes individual name, race, and country of origin
- Target groups: Chinese and Japanese (enumerated as separate ethnic groups)
- Only coarse race-geography information (i.e., Asian) is available
- Two methods: oracle ℓ BISG and coarse prior ℓ BISG
- Ask Gemini 3.0 Flash to propose three short descriptions that identify each subgroup
 - Cantonese families from the Pearl River Delta settling in California and the American West
 - Taishanese and Sze Yup laborers migrating to the sugar plantations of the Hawaiian Islands
 - Fujianese and Hokkien merchants and sailors establishing communities in Atlantic port cities like New York
- Same scripts for the other racial groups while adding “in the United States around 1900”
- Outcome: the top quartile of the 1900 socioeconomic index within each county

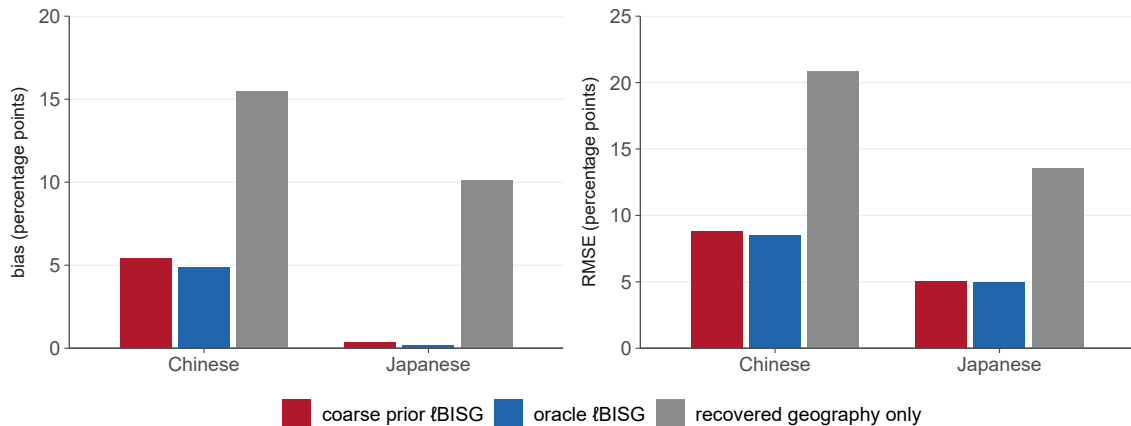
Coarse prior ℓ BISG is as Accurate as Oracle ℓ BISG



Coarse prior ℓ BISG is as Calibrated as Oracle ℓ BISG



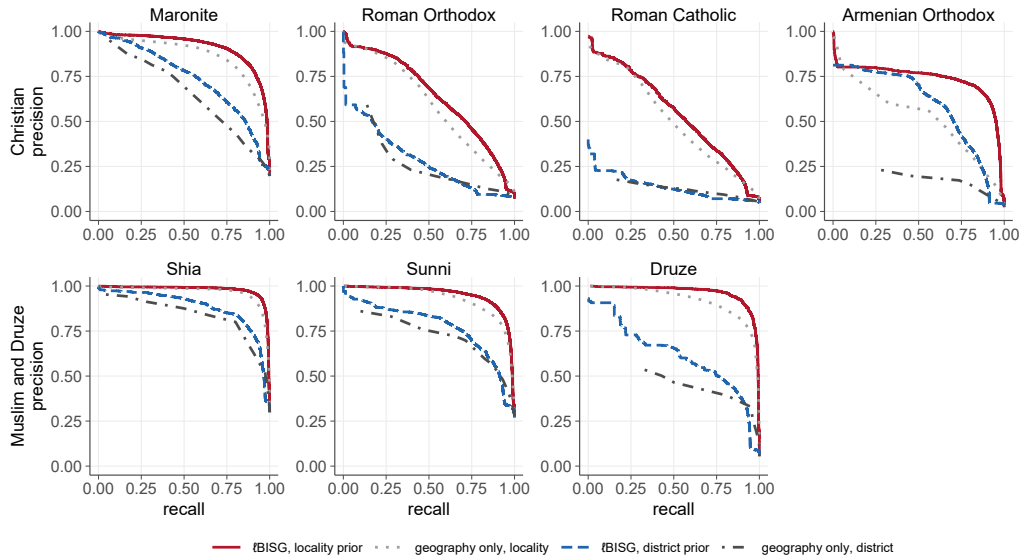
Downstream Analysis Performs Well



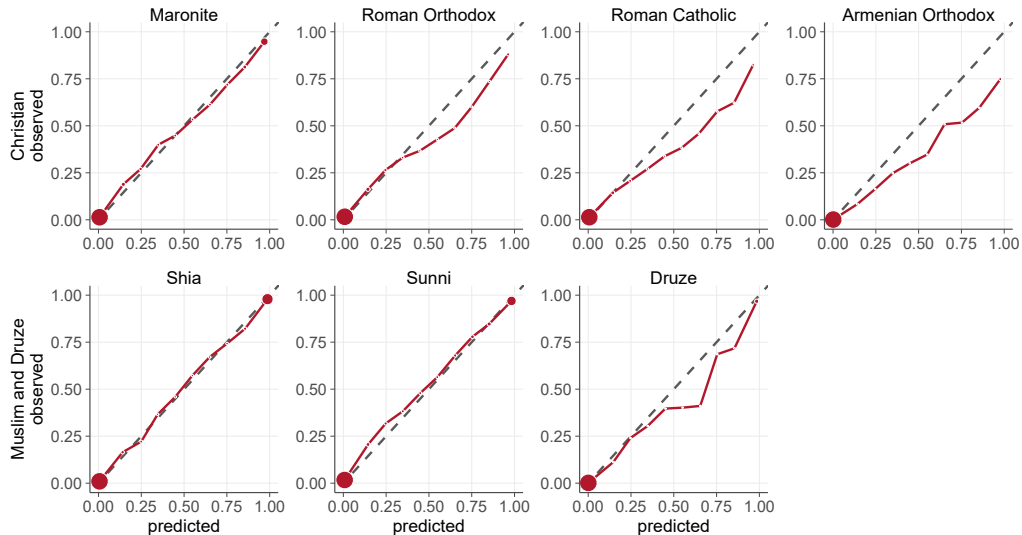
Validation III: Applying ℓ BISG Outside of the United States

- Lebanon voter file published in 2018 by the Ministry of Interior and Municipalities
 - Name, residence, self-reported religious sect
 - Shia (29.9%), Sunni (27.1%), Maronite (19.8%), Roman Orthodox (7.0%), Druze (5.5%), Roman Catholic (4.9%), and Armenian Orthodox (2.4%)
 - Highly segregated at localities but not at districts
-
- Use Gemini 3.0 Flash to create lists via prompts written in Arabic
 - Embed surnames using the paraphrase-multilingual-mpnet-base-v2 model, which accommodates Arabic script
 - Outcome: the share of each sect born in 1985 or later within each district

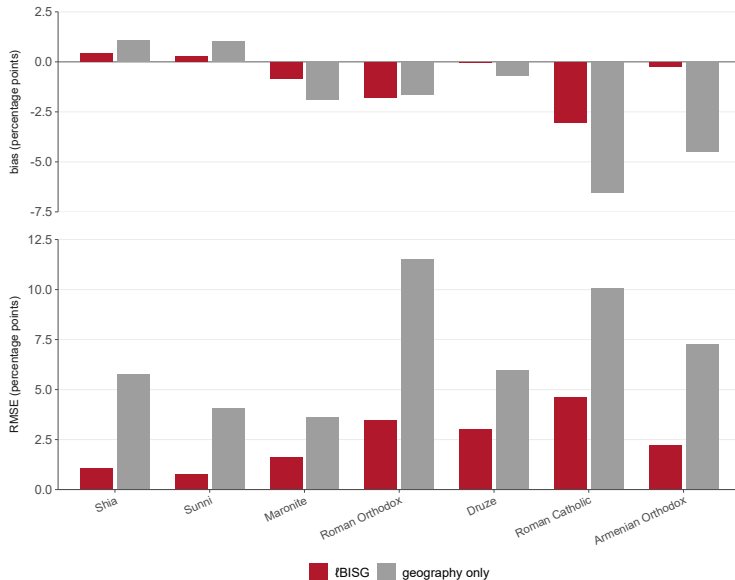
ℓ BISG Predicts Religious Sectors Accurately



*l*BISG Probabilities are Calibrated



ℓ BISG Downstream Analysis Is Accurate



Concluding Remarks

- Downstream analysis with predicted race probabilities
 - standard methods (weighting and thresholding) are likely to be biased
 - use BIRDIE instead: name as a proxy for race
 - debiased machine learning to account for estimation error of clusters
 - a similar issue arises in statistical analysis with ML prediction
- Improving racial prediction
 - standard BISG works well but is not widely applicable
 - ℓ BISG uses name lists for probabilistic prediction
 - ℓ BISG produces accurate and calibrated race probabilities in three empirical validations
- Open-source software implementation
 - **wru**: <https://cran.r-project.org/package=wru>
 - **BIRDIE**: <https://cran.r-project.org/package=birdie>

