FCSM 2022:
Estimating Missing Race and Ethnicity Data with Surname and Geolocation

Jeff Tessin

Assistant Director (Statistics)
Applied Research and Methods Group
U.S. Government Accountability Office (GAO)
Washington, DC
tessinj@gao.gov | 202-512-4232

October 2022

The Problem

How do we analyze how federal programs work across racial and ethnic groups, when agencies do not measure race and ethnicity?

- Race/ethnicity is not explicitly needed to administer many programs. Measurement may enable or suggest discrimination.
- Yet, race/ethnicity often matters. Race-blind administration is not necessarily equitable.

The Problem

Many researchers and policymakers grapple with a lack of race and ethnicity in administrative data.

- Health disparity research: COVID-19 cases, cardiovascular health outcomes, quality of health care
- Voter registration and voting activity

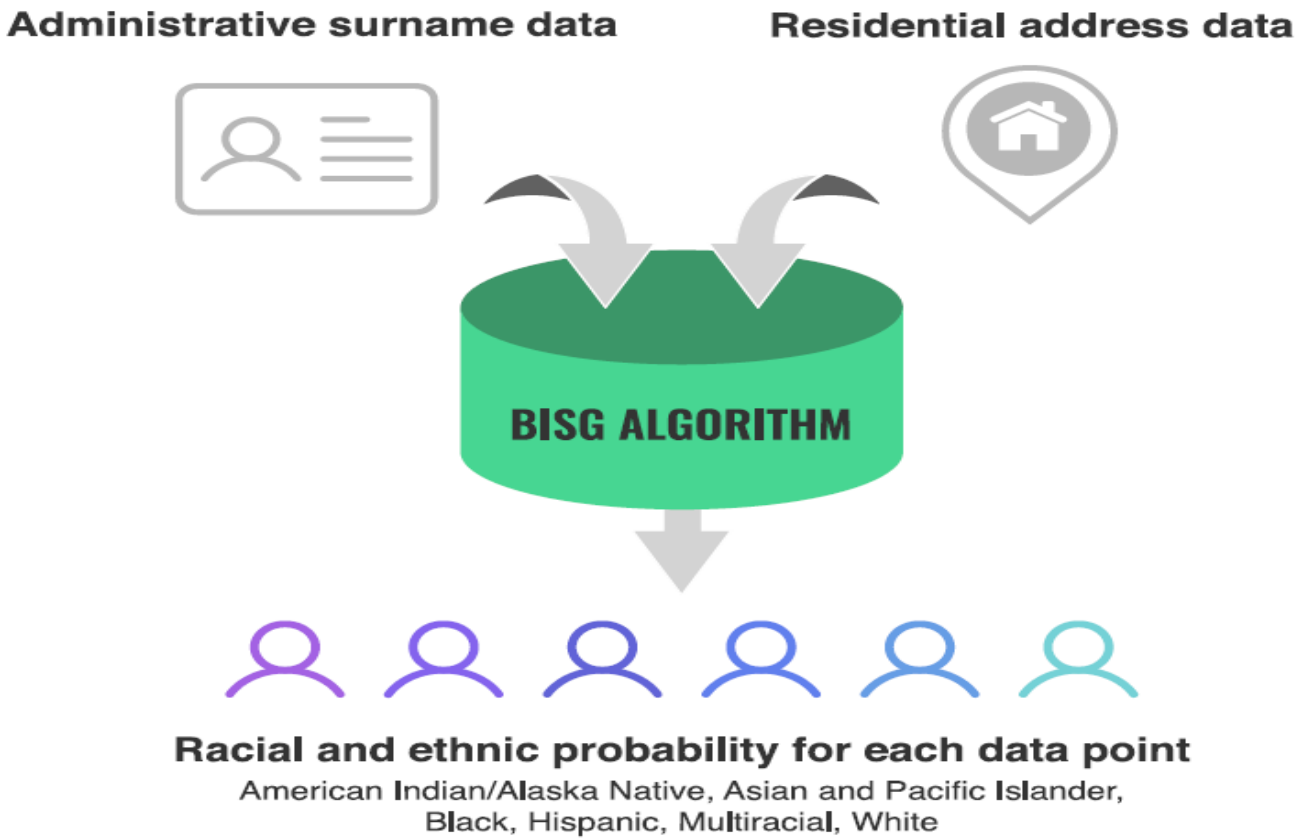
3 Types of Solutions

- Direct measurement
- Linkage
- Imputation

Lit Review: Imputation Methods

- **Literature Search Parameters:**
  - Material types: scholarly/peer reviewed material, conference papers, working papers, nonprofit/think tank publications
  - Publication date range: 2006 - 2021
  - Databases searched: Scopus, ProQuest, EBSCO, Harvard Think Tank
- **Imputation Methods Identified by Literature:**
  - Statistical modeling (e.g., GLMs)
  - Data mining and machine learning algorithms
  - Probabilistic classifiers
  - Bayesian Improved Surname Geocoding

BISG: Bayesian Improved Surname Geocoding



GAO

BISG: Bayesian Improved Surname Geocoding

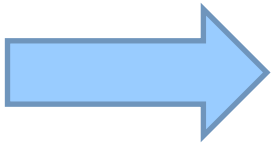
- Has been applied to financial, healthcare, voter registration, marriage license datasets
- Used by the Federal Reserve, Consumer Financial Protection Bureau, Medicare/Medicaid
- Needs only two commonly measured variables to impute with reasonable accuracy

Source: RAND (link)

BISG: Bayesian Improved Surname Geocoding

Auxiliary Dataset		
Person	**Observed Race**	**Predictor**
1	Asian	Group 1
2	Asian	Group 1
3	Asian	Group 1
4	Black	Group 2
5	Black	Group 2

	Asian	Black	Row Total
Group 1 (Row pct)	3 (100%)	0 (0%)	3
Group 2 (Row pct)	0 (0%)	2 (100%)	2



Target Dataset			
Person	**Imputed Probability Asian**	**Imputed Probability Black**	**Predictor**
1	100%	0%	Group 1
2	0%	100%	Group 2

GAO

BISG: Bayesian Improved Surname Geocoding

Many surnames can predict race/ethnicity with minimal error

Surname	Probability Multiracial	% American Indian/ Alaskan Native	% Asian / Pacific Islander	% Black	% Hispanic	% White	% Other
SANTUARIO	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
HERNANDEZA	0.0%	0.0%	0.0%	0.3%	99.4%	0.3%	0.0%
ZHEN	0.5%	0.0%	98.6%	0.0%	0.3%	0.6%	0.5%
ZHOU	0.5%	0.1%	98.2%	0.2%	0.1%	1.0%	0.6%

GAO

BISG: Bayesian Improved Surname Geocoding

Some surnames are less predictive (more equally represented among each group)

Top 5 Names in the 2010 Census, by Race

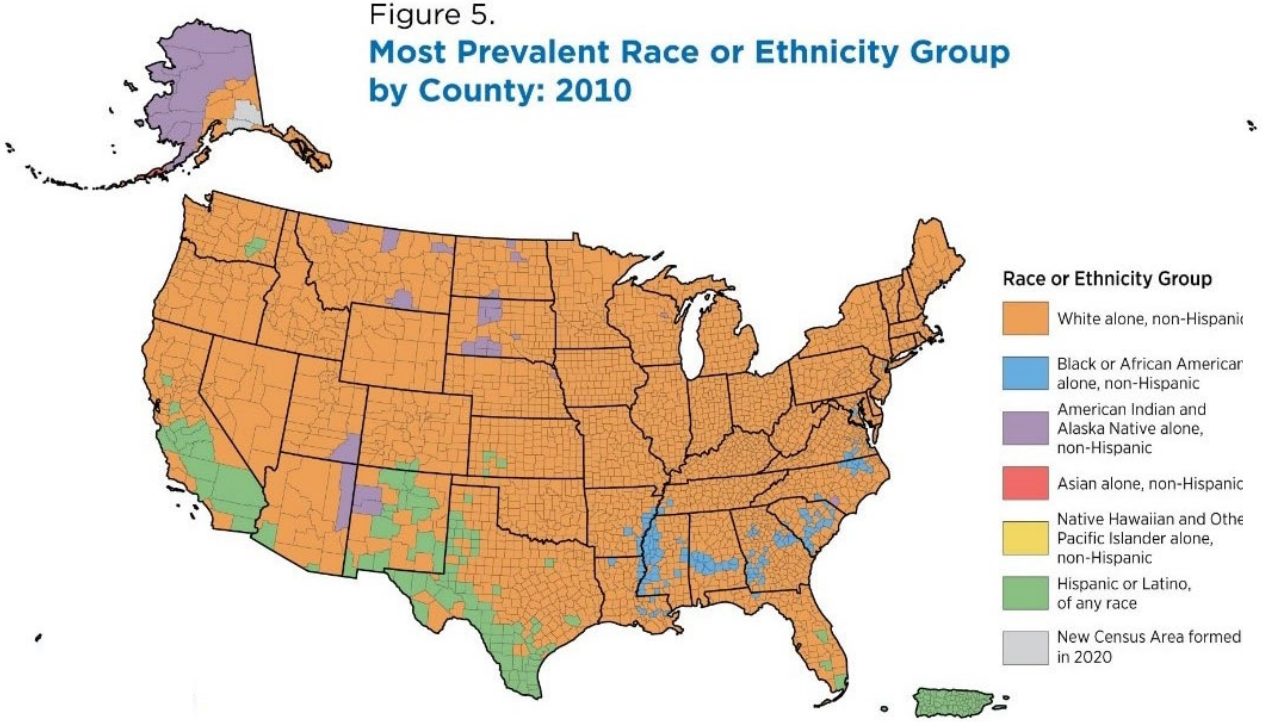
Surname	Probability Multiracial	% American Indian/ Alaskan Native	% Asian / Pacific Islander	% Black	% Any Hispanic	% White	% Other
SMITH	2.2%	0.9%	0.5%	23.1%	2.4%	70.9%	3.1%
JOHNSON	2.6%	0.9%	0.5%	34.6%	2.4%	59.0%	3.5%
WILLIAMS	2.8%	0.8%	0.5%	47.7%	2.5%	45.8%	3.6%
BROWN	2.6%	0.9%	0.5%	35.6%	2.5%	58.0%	3.4%
JONES	2.6%	1.0%	0.4%	38.5%	2.3%	55.2%	3.6%

Source: 2010 Census Surname Table (https://www.census.gov/topics/population/genealogy/data/2010_surnames.html)

BISG: Bayesian Improved Surname Geocoding

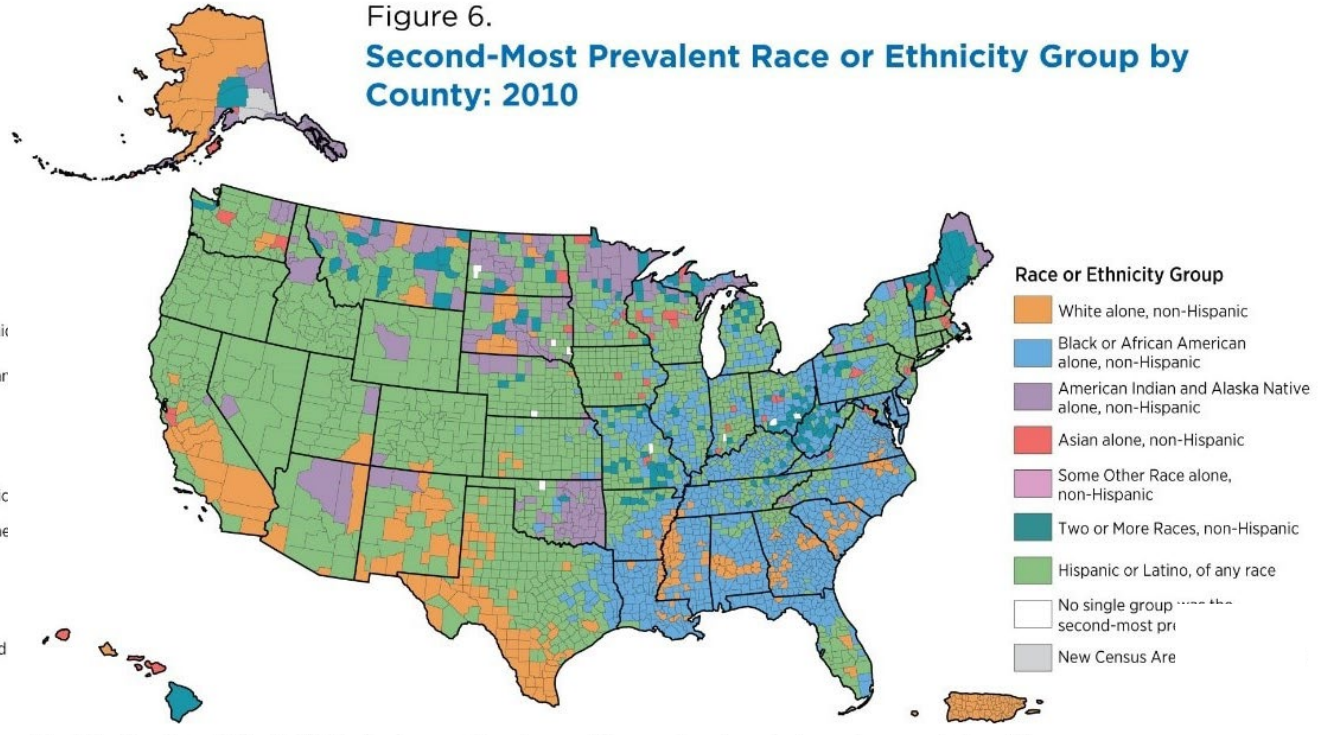
Adding residential geolocation improves predictions, due to clustering among some groups with less predictive names

Figure 5.
Most Prevalent Race or Ethnicity Group by County: 2010



Note: Some Other Race alone, non-Hispanic and Two or More Races, non-Hispanic were not the most prevalent group in any county. Native Hawaiian and Other Pacific Islander, non-Hispanic was the most common group in Kalawao County, HI.
Source: U.S. Census Bureau, 2010 Census Redistricting Data (Public Law 94-171) Summary File.

Figure 6.
Second-Most Prevalent Race or Ethnicity Group by County: 2010



Note: Native Hawaiian and Other Pacific Islander alone, non-Hispanic was not the second-most prevalent group in any county. Some Other Race alone, non-Hispanic was the second-most common group in Dukes County, MA.
Source: U.S. Census Bureau, 2010 Census Redistricting Data (Public Law 94-171) Summary File.

Source: Census (link)

GAO

BISG: Bayesian Improved Surname Geocoding

- Estimate by combining surname and geolocation probabilities from 2010 Census, using Bayes Rule:
  - *G* = {block groups}
  - *R* = {racial/ethnic groups}
  - *S* = {surnames}
  - *i* = {1, 2, …, *N* people}

$$\Pr(G_i = g|R_i = r) = \frac{\Pr(R_i = r|G_i = g) \cdot \Pr(G_i = g)}{\sum_g \Pr(R_i = r|G_i = g) \cdot \Pr(G_i = g)} \quad (1)$$

$$P_{ir} = \Pr(R_i = r|S_i = s, G_i = g) = \frac{\Pr(G_i = g|R_i = r) \cdot \Pr(R_i = r|S_i = s)}{\sum_r \Pr(G_i = g|R_i = r) \cdot \Pr(R_i = r|S_i = s)} \quad (2)$$

GAO

BISG: Bayesian Improved Surname Geocoding

- Validate predictions against self-reported race/ethnicity
  - 2010 Census
  - Mortgage applications
  - Custom surveys (healthcare)

- Generally high probability/certainty, though less accurate for Blacks and Whites than Asians and Hispanics

GAO

BISG: Bayesian Improved Surname Geocoding

Table 5 Percentage of individuals with specified BISG probabilities, for each of the six predicted racial/ethnic categories

Bayesian probability	Hispanic	Asian	Black	AI/AN	Multiracial	White
0 to <0.05	87.9	93.0	83.1	100.0	100.0	4.3
0.05 to <0.20	2.5	2.6	8.6	0.0	0.0	11.2
0.20 to <0.50	0.9	0.7	2.6	0.0	0.0	3.6
0.50 to <0.90	6.2	0.7	2.9	0.0	0.0	12.6
0.90 to 1	2.5	2.9	2.8	0.0	0.0	68.3

Source: Elliott, et al. 2009

BISG: Application to COVID Tax Credits

- What options are available to examine disparities in the proportion of businesses benefiting from selected tax provisions by the sex, race, or ethnicity of the business owner? (GAO-22-104582)

- Scope: CARES Act tax relief provisions for employers and self-employed business owners; businesses of interest were small, single-owner firms

BISG: Application to COVID Tax Credits

- Problem: IRS doesn't measure race/ethnicity!
  - According to IRS officials, race/ethnicity is not collected because it is not needed for administration of tax code.
  - Treasury recently announced that they will begin equity analyses of tax policies.

GAO

BISG: Application to COVID Tax Credits

- Two input data files from 2010 Census (most recent):
  - Summary File 1: race distribution by block group
  - Surname Table: race distribution by surname

- Extracted surnames and addresses from various tax filings
  - Geocoded addresses into block groups using default SAS process (PROC GEOCODE). Provides a useable Census block group for roughly 85% of available addresses

GAO

BISG: Application to COVID Tax Credits

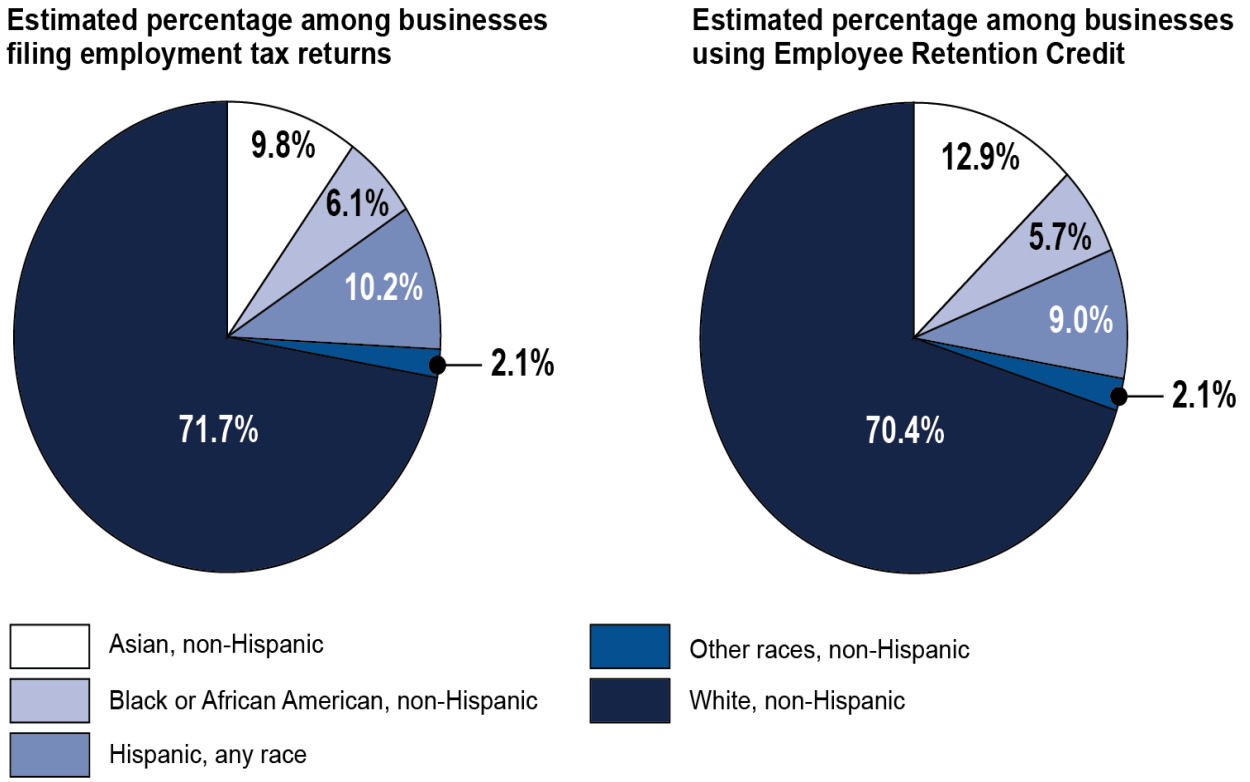
- Due to measurement constraints, collapsed groups into: Hispanic and Non-Hispanic Asian, Black, White, and Other
- Estimation involves algebra and can be implemented in various software
- No validation data

GAO

BISG: Application to COVID Tax Credits

- Imputed race/ethnicity using best available combination of name and location.
  - High rate of matching to surname and Census block group location, but not 100%
  - When location was unavailable, used name alone
  - When names don’t appear in Census data, use probabilities for a residual “All Other Names” group.

BISG: Application to COVID Tax Credits



Source: GAO analysis of Internal Revenue Service taxpayer data, Social Security Administration data, and U.S. Census Bureau data. | GAO-22-104582

GAO

BISG: Application to Flood Insurance Data

- Ongoing work analyzing how revised flood risk rating system affects FEMA policyholders by race and ethnicity
- Policyholder data have surname and address
- BISG estimates feasible, largely due to reliable block groups
- Surveys, linkage too expensive or infeasble

GAO

BISG: Application to SSA Data

- Evaluating the quality of SSA service delivery during COVID pandemic, especially by race and ethnicity
- Analyzing self-reported data from SSA:
  - Missing for many applicants
  - Before 2009: measured “Black,” “White,” and “Other” via paper form
  - After 2009: measured OMB categories

GAO

BISG: Application to SSA Data

- Experimenting with BISG, applied to benefit claims
  - Surnames and addresses available from SSA
  - Uncertain feasibility and accuracy
  - Possible error from surrounding block group, due to institutional living

Implications

- Greater interest in measuring race/ethnicity among program participants

- Imputation can enrich many administrative datasets with limited direct measurement.

GAO

Implications

- BISG predicts with minimal error, using only two variables that agencies often do collect reliably.

- Enables analysis of race/ethnicity when direct measurement is not desirable, possible, or affordable

GAO

Implications

- Estimates vary in quality. Imputation error exists.
  - Can reduce risk by using only the estimates with high confidence
  - Sensitivity analysis
  - Validate with self-reported data, when possible
- Must assume that responses to the decennial census resemble responses that specific populations would give.

GAO

Implications

- Can’t measure small or detailed racial/ethnic groups (e.g., Irish Catholic Whites, Multiracial Hispanics)
- Latest Census data (2010) are old. Need updates and re-validation.
- Predictive power may decrease with more nuanced identification patterns and less residential segregation.

GAO

Resources

- Elliott, et al., “Using the Census Bureau’s Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities,” *Health Services, Outcomes, and Research Methods* (2009) 9: 69-83.
- Haas, et al., “Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity,” *Health Services Research (*2019) 54:13–23.

Thanks for listening!