FAIR

FAIR AND RESPONSIBLE BANKING FORUM

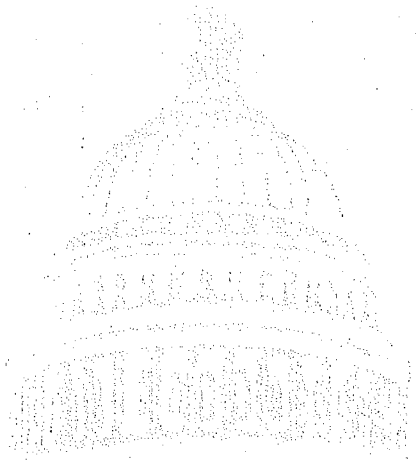
CBA LIVE 2014

RED, WHITE + BANKING

WASHINGTON, DC | MARCH 31–APRIL 2

Presented by:

Bernard R. Siskin, Ph.D.
Director, BLDS



EMERGING TRENDS IN QUANTITATIVE ANALYTICS

FAIR

Estimation of Unknown Race

- Overview of alternative methodologies
- BISG: Bayesian Improved Surname Geocoding: CFPB Choice for Race
- What is it?
- How is it computed?
- What is it used for?
- How is it used?
- How good is it? What are its problems? What are its issues?
- Gender Proxy?

FAIR

Overview of Alternatives: Race/Ethnicity

- Name Recognition: Surname
  - Works well for Hispanics and Asians, limited utility for African Americans and whites
- Geocoding:
  - Works well in highly segregated areas where many individuals of a particular race are heavily concentrated. Used best at census block or block group level.
  - Accepted by Courts
- Third Party Sources:
  - Can be good, but not readily available, nor transparent
- BISG: Combines surname and geocoding to achieve best generally available estimate of race

FAIR

Overview of Key Assumptions of Methodologies

Geocoding:

Within the defined geographic area, the probability of "selection" is independent of race

Name Recognition:

Given your surname, the probability of "selection" is independent of race

BISG:

Given your surname, the probability of "selection" is independent of race, plus given your race, the probability you would live in a defined area is independent of your name

FAIR

BISG: What is it?

- A statistical method of estimating the probability someone is of one of 6 races:
  - White non-Hispanic
  - Hispanic
  - African American non-Hispanic
  - Asian Pacific Islander
  - American Indian/Alaska Native
  - Multi-race

BISG: How Is It Computed?

- BISG combines surname information with geocoding information via Bayesian Formula
- Starts with name probability, then adjusts based on geocoding data
- "Hispanic" category includes all Hispanics, while all other races (including African Americans) are non-Hispanic

ISSUES:

- How geographic area should be defined, and what controls (i.e., age, home ownership, etc.) should be used to improve definition of geographic area

FAIR

BISG: What Is It Used For?

Classification

- Classifies individuals as white, African American, etc. Then uses those classified as the population to study. For this population, race is considered to be known.

Statistical Estimates

- Estimates frequencies (not specific individuals) by race. Used for estimating disparate impact (differences in outcomes by race)
- Class-wide "damages"
- Total number of (not individual) "victims"

FAIR

BISG: How Is it Used?

- Classification: Assigns each observation to a race, or leaves as "unknown" based on BISG probabilities

- Proportional Estimation: Uses BISG probabilities directly to estimate counts or do statistical estimates of race effect, but does not assign a specific race to an individual

FAIR

BISG: Classification

- 80 Percent Rule Industry Standard
  - e.g., if probability of race ≥ 0.80, assign to that race; if no race has probability ≥ 0.80, assign as "unknown"
- 90/80 white if $P_w \geq 0.90$
  - Specific minority if $P_m \geq 0.80$; else, unknown
- 90 or 80/MAX
  - White if $P_w \geq 0.90$ or 0.80; if $P_w < 0.5$ assign to minority with highest P value; if $P_w < 0.90$ or 0.80, but $P_w \geq 0.50$ assign as unknown.
- MAX
  - assign to race with maximum probability

FAIR

How Good Are BISG Classification Estimates?

- Coverage of minorities is often poor.
- If used to identify "victims" : there will be some false positives and process may miss many victims. Coverage can be increased by use of the 50 Percent/Max Rule, but that will also increase false positives.
- Good if used to estimate racial differences or matched pairs since random errors mask impact and do not create impact.
- However, due to omitted variable bias, the adverse impact based on studying only segregated areas is greater than the adverse impact observed when also considering mixed areas. Since classification does not study mixed areas, it overstates the true disparate impact.

FAIR

BISG: How Is it Used?
Proportional Estimation

- Does not identify individuals. Simply uses probabilities to do statistical analyses
- Proportional estimation is based on the assumption of proportionality in a population: i.e., based on the assumption that if, for example, 100 individuals have a 0.10 probability of being African American, then 10 of the 100 are African American

					Proportional Estimation			
	BISG Probability		Number		African American		White	
Applicants	African American	White	Accepted	Rejected	Accept	Reject	Accept	Reject
100	0.10	0.90	90	10	9	1	81	9
100	0.50	0.50	50	50	25	25	25	25
100	0.90	0.10	10	90	9	81	1	9
		Proportional Estimate of Impact			43	107	107	43
					28.70%		42.70%	
		Classification Estimate of Impact			10	90	90	10
					10.0%		90.0%	

FAIR

BISG: How Is It Used?
Proportional Estimation (Continued)

- The proportional estimation methodology just discussed assumes disparate impact only (not disparate treatment). Individuals with the same probability of being African American (i.e., primarily in the same geographic area in which economic factors are relatively the same) are assumed to be treated the same

BISG: How Is It Used? Proportional Estimation (Continued)

- Use of BISG probabilities directly in regression. Race coefficients interpreted as race effects. Race effect is thus estimate of difference in outcome of observations with 100 percent race BISG versus 100 white non-Hispanic.
- Measures disparate treatment effect. Valid to test for statistical evidence of impact or treatment discrimination but overstates effect if disparate impact exists (alone or in combination with disparate treatment).
- Alternative is "weighted" regression. Limitation is that it assumes only disparate impact and, hence, may understate true effect if disparate treatment exists.

FAIR

How good are BISG Proportional Estimates?

Proportional Estimation

- Very good proxy. All statistical literature supports that it is the best generally available statistical methodology.
- Reliable and accurate for statistical studies.
- Uses all the data.

Issue

- How to reconcile when both disparate treatment and disparate impact are present.

FAIR

Gender Proxy

- First name used to assign probabilities
- Two sources:
  - 1990 Census tabulation of first names by gender for all names to cover at least 95 percent of population.
  - Since 1880 the Social Security Administration has recorded all first names of newborns with an occurrence rate of at least 10 per year.
- There are generally few "unknowns" if the classification methodology uses a "greater than 20 percent, but less than 80 percent" standard to determine the probability of being female.
- "Unknowns" occur when there is no name match. The use of social security number typically results in a small number of "unknowns".

FAIR

APPENDIX

Bayesian Improved Surname Geocoding (BISG) Calculation Quick Start

The BISG is calculated in the following manner.

$$P_{BISG} = \frac{P_{Nm_i} * P_{Geo_i}}{\sum_j P_{Nm_j} * P_{Geo_j}}$$

Where *i* is the race of interest and *j* is each of the race/ethnicities covered by the surname table (e.g. White, Black, Asian/PI, Native American, Hispanic, Multi-Racial).

P_{Nm} is obtained for each surname from a table which has been derived from the frequently occurring Census 2000 names (http://www.census.gov/genealogy/www/data/2000surnames/). That data has been modified slightly to ensure that the surname racial probabilities always sum to 1.0. If a surname is not covered by the census table, default population proportions should be substituted. The default proportions are:

White:	0.6938	Asian/PI:	0.0689
African American:	0.1112	Multi Racial:	0.0079
Native American:	0.0089	Hispanic:	0.1093

P_{Geo} is calculated for each tract/blockgroup/block, depending on the quality of your geocoding, as the ratio of the number of people of a specified race in the geocoding area divided by the national total of people of that race. Please note variables with 18 in the name are 18+ only. Variables with HISP indicate Hispanic individuals, NHISP indiciates Non-Hispanic individuals. SUMLEV indicates the Level of Geographic Summary where 140 = Tract, 150 = Block Group and 750 = Block. The relevant geocoding fields and population totals are :

White:	Variable = WHITE18_NHISP; Total = 157,123,289
Black :	Variable = BLACK18_NHISP; Total = 27,327,470
Asian/PI:	Variable = ASIAN18_NHISP+HAWAII18_NHISP; Total = 11,637,514
Native American:	Variable = AMIND18_NHISP; Total = 1,600,043
Multi Racial:	Variable = TWORACE18_NHISP; Total = 3,177,961
Hispanic:	Variable = TOTAL18_HISP; Total = 36,138,485

FAIR

Bayesian Improved Surname Geocoding (BISG) Calculation Quick Start (Continued)

- **Recommendations for Zip Code Based Geocoding:** If possible, geocoding using a Zip+4 can be accurate to the block level given the small area encompassed by Zip+4. If an address cannot be geolocated using a Zip+4, then full address geocoding is recommended.

- **Recommendation for Zero Population Areas:** It is possible that a census tabulation area may have had zero occupants at the time of the data collection. The BISG requires non-zero geocoding populations, so it is recommended that the smallest geographic area is used based on the quality of your geocoding, which is escalated to larger areas when the population is zero. The correct geographic area should be the smallest of the block/block group/tract populations which has a non-zero population.

- **Example Calculation of BISG:**
- **Last Name:** Smith
- **Surname Racial Proportions**
  - White = 0.7334
  - Black = 0.2222
  - Asian/Pacific Islander = 0.0040
  - American Indian/Alaskan Native = 0.0085
  - Two or More Races: 0.0163
  - Hispanic: 0.0156

- **Zip Code:** 01001-1648
- **Census Geocoding Population (Proportion of Total Population)**
  - White = 470 (470/157123289 = 2.99128E-06)
  - Black = 14 (14/27327470 = 5.12305E-07)
  - Asian/Pacific Islander = 13 (13/11637514 = 1.11708E-06)
  - American Indian/Alaskan Native = 0 (0/1600043 = 0)
  - Two or More Races: 3 (3/3177961 = 9.44002E-07)
  - Hispanic: 22 (22/36138485 = 6.08769E-07)

FAIR

Bayesian Improved Surname Geocoding (BISG) Calculation Quick Start (Continued)

- BISG Calculation
- Denominator (Constant for all Races in this BISG Calculation)
- $(P_{Nm_W} * P_{Geo_W})+ (P_{Nm_B} * P_{Geo_B})+(P_{Nm_{As}} * P_{Geo_{As}})+ (P_{Nm_{AI}} * P_{Geo_{AI}})+ (P_{Nm_{2race}} * P_{Geo_{2race}})+ (P_{Nm_{Hisp}} * P_{Geo_{Hisp}})$
- (0.7334*2.99128E-06)+(0.2222*5.12305E-07)+(0.004*4.46831E-09)+(0.0085*0)+(0.0163*9.44002E-07)+(0.0156*6.08769E-07)
- 2.19381E-06+1.13834E-07+4.46831E-09+0+1.53872E-08+9.4968E-09
- 2.33699E-06
-
- BISG Race Proportion

• White:	2.19381E-06/2.33699E-06	= 0.938730435
• Black:	1.13834E-07/2.33699E-06	= 0.048709689
• Asian/Pacific Islander:	4.46831E-09/2.33699E-06	= 0.001911991
• American Indian/Alaskan Native:	0.00000E00/2.33699E-06	= 0.000000000
• Two or More Races:	1.53872E-08/2.33699E-06	= 0.006584200
• Hispanic:	9.49680E-09/2.33699E-06	= 0.004063685

FAIR

References

- Elliott MN, Morrison PA, Fremont A, McCaffrey DM, 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.

- Elliott MN, Fremont A, et al. (2008). "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." *Health Services Research*, 43(5pl): 1722-1736.

- McCaffrey D & Elliott MN. (2008). "Power of Tests for a Dichotomous Independent Variable Measured with Error." *Health Services Research*, 43(3): 1085-1101.

FAIR