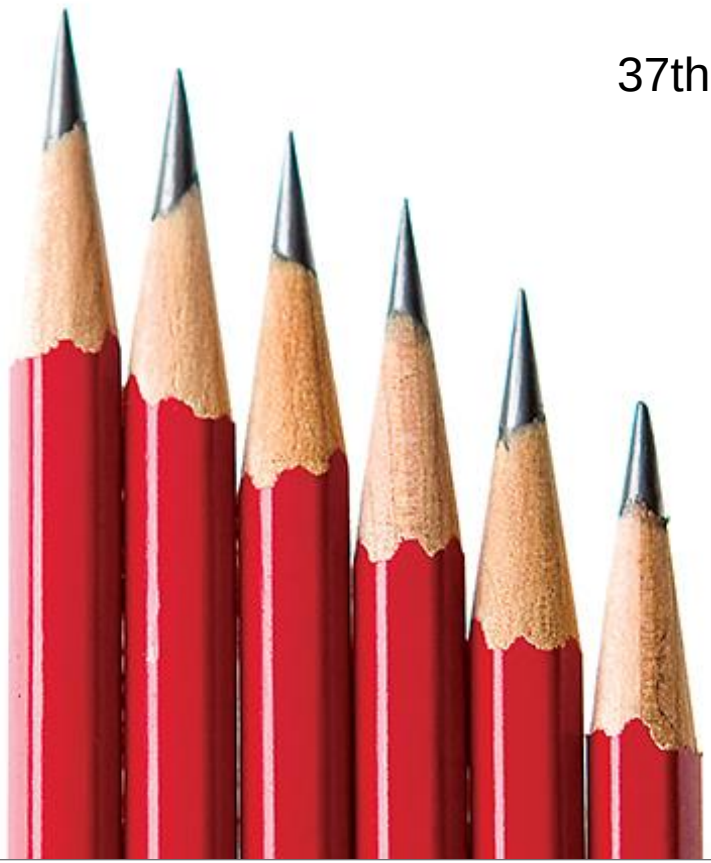
Proxy Methods in Fair Lending Analysis



California Bankers Association
37th Annual Regulatory Compliance Conference
October 8, 2015

David Skanderson, Ph.D.
Vice President
dskanderson@crai.com
202.662.3955

CRA Charles River Associates

This presentation is not complete without the accompanying discussion.

Agenda

- Why use proxies?
- Proxies from the past
- New kid on the block: BISG
- Practical application
- Are proxy-based methods sufficiently reliable?
- Unresolved issues

Why proxy-based methods are used

- Anti-discrimination laws apply to a broad array of lending, but …
- … self-reported race, ethnicity, gender are not available outside of HMDA-reportable home mortgage lending
  - Home equity lending (non-HMDA)
  - Automobiles, motorcycles, recreational vehicles
  - Credit cards
  - Unsecured personal lending
  - Payday lending
  - Small business & commercial lending
  - Etc.

"Traditional" proxy methods

- **Race/ethnicity:** Either surname or geography
  - Surname
    - Works best for names strongly associated with a particular race/ethnicity
    - Works better for Hispanics & Asians than for other minorities
    - Works very poorly for African Americans
  - Geography/neighborhood
    - Works better for individuals in highly segregated areas
    - Therefore, tends to work better for African Americans than for other groups
- **Gender:** First name

New kid on the block: "BISG"

- *"Bayesian Improved Surname Geocoding"*
- Originally used in health care research
- Combines surname and geographic information
- Attempts to reduce the shortcomings of using the either piece of information in isolation
- A *probabilistic estimate of* a consumer's race/ethnicity characteristics

Source: "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," M. Elliot, et al. (2009)

How BISG proxies are calculated

- Derives a set of race/ethnicity probabilities for each consumer
  - Starts with surname-based probabilities
  - Adjusts them using geography-based demographic data
  - Represents an estimate of the probability that a person is of a particular race, given the person's surname and the demographic composition of their neighborhood

How BISG proxies are calculated

Step 1: Determine surname probabilities

Race/Ethnicity Probabilities for surname "Johnson"	
Race/Ethnicity	**Share**
Hispanic	1.5%
African American	33.8%
Asian/PI	0.4%
American Indian	0.9%
White	61.6%
2+ Races	1.8%
Total	100.0%

Source: Census Bureau

How BISG proxies are calculated

Step 2: Determine geography-based demographics

18+ Population of Tract 0050.02 - Washington, DC				
Race/Ethnicity	**Tract Counts**	**Intra-Tract Shares***	**U.S. 18+ Population Count**	**Share of U.S.**
Hispanic	1,340	24.5%	36,138,485	0.0037%
African American	1,008	18.4%	27,327,470	0.0037%
Asian/PI	307	5.6%	11,637,514	0.0026%
American Indian	15	0.3%	1,600,043	0.0009%
White	2,693	49.2%	157,123,289	0.0017%
2+ Races	109	2.0%	3,177,961	0.0034%
Total	5,472	100.0%	237,004,762	0.0023%

Source: Census Bureau

*Important – BISG does not use the Intra-tract shares commonly used in other geography-based proxies.

How BISG proxies are calculated

Step 3: Calculate BISG Probabilities

BISG Calculation Example			
Race/Ethnicity	Surname "Johnson"	Tract 0050.02 Wash, DC	BISG Probability Vector
Hispanic	1.5%	0.0037%	2.3%
African American	33.8%	0.0037%	51.1%
Asian/PI	0.4%	0.0026%	0.5%
American Indian	0.9%	0.0009%	0.3%
White	61.6%	0.0017%	43.2%
2+ Races	1.8%	0.0034%	2.6%
Total	100.0%	0.0023%	100.0%

Source: Census Bureau

Let x_1 be the product of the surname probability and the tract percentage for race/ethnicity Group 1.

BISG probability for Group 1 = $x_1 / (x_1 + x_2 + x_3 + x_4 + x_5 + x_6)$

According to BISG, my next-door neighbors are …

		My African American Neighbor		My Non-Hispanic White Neighbor	
Race/Ethnicity	Block Group Race/ Ethncity Shares	Surname Probabilities	**BISG Probability Vector**	Surname Probabilities	**BISG Probability Vector**
Hispanic	20.1%	1.4%	**1.7%**	1.6%	**1.5%**
White	41.2%	76.3%	**43.1%**	60.7%	**26.1%**
Black/African American	33.6%	19.7%	**52.2%**	34.5%	**69.7%**
Native American	0.0%	0.6%	**0.0%**	0.8%	**0.0%**
Asian/Pacific Islander	2.6%	0.3%	**0.2%**	0.4%	**0.1%**
Other/Multi-Race	2.6%	1.7%	**2.9%**	1.9%	**2.5%**
Total	100.0%	100.0%	**100.0%**	100.0%	**100.0%**

The devil is in the details

- How to parse applicant surname data for matching to Census surname list?
  - Hyphenated names (e.g., are LOPEZ-GARCIA, LOPEZ GARCIA and LOPEZGARCIA the same name?)
  - Accent marks, other punctuation, extraneous spaces
  - Initials & suffixes (Jr., Sr., II, III, IV, MD, DDS, PhD)
  - Business names
  - Transposed names
  - Etc.
- How to assign race/ethnicity for hyphenated or compound surnames (e.g., Angelina Jolie-Pitt, Kim Kardashian West)?
  - CFPB's approach does not appear to be consistent with methods Census Bureau used in constructing the name list

The devil is in the details

- How to treat records with unmatched surnames
- Choice of geographic level: block, block group, tract or ZIP demographics?
- How to treat addresses without an exact geocode match at the chosen geographic level?
- Total population demographics or adult (18+) population demographics?

The devil is in the details

- How to treat consumers whose surnames are not on the Census list?
- How to deal with surname probabilities that do sum to 1.0 across all race groups (due to redacted probabilities), or that sum to more than 1.0 (due to rounding)?
- How to assign BISG probabilities to loan records with multiple applicants – which applicant to use for coding?

Using proxies in fair lending analysis: Threshold approach

- Discrete probability threshold
  - Example: Assume Hispanic ethnicity if BISG Hispanic probability is 80% or higher
- Sensitivity analysis based on alternative thresholds
- Advantages:
  - Straightforward to apply & interpret in regression models and other monitoring
  - Straightforward to identify “probable minorities” for file review sampling
- Disadvantage:
  - Tends to exclude a large proportion of the data sample

Using proxies in fair lending analysis: CFPB Approach

- Proportional Race/Ethnicity Assignment
  - Loan records not assigned to a definitive group
  - Each consumer is assumed to be a composite of up to six race/ethnicity groups: X% white, Y% Hispanic, Z% African American, etc.
  - Essentially, each loan record is weighted based on its BISG probabilities
  - All consumers are aggregated to draw conclusions about racial disparities in the combined consumer sample

(Also known as "continuous" or "spectrum" approach)

Using proxies in fair lending analysis: CFPB Approach

- Advantage:
  - Minimize number of data records excluded from analysis
- Disadvantages:
  - More difficult to implement and interpret than discrete threshold
  - Extremely difficult to interpret results of underwriting analysis (logistic regression)

Note: CFPB has not published its analysis methods, except for proxy derivation.

Regression model under proportional approach

- Example: Regress auto dealer mark-up on BISG probabilities

$$\text{Mark-up} = a + b_{AA}*P_{AA} + b_{H}*P_{H} + b_{API}*b_{API} + b_{NA}*P_{NA} + b_{MR}*P_{MR}$$

- The non-Hispanic white probability is omitted, so that all race effects are measure relative to non-Hispanic whites

- b_i = the change in dealer mark-up associated with a one percentage point increase in the probability of being race *i* (i.e., the disparity estimate)

CFPB approach to estimating consumer harm in auto lending cases

- Step 1: Estimate disparity
- Step 2: Estimate aggregate $ harm over entire data sample based on disparity estimate (including all loan records regardless of BISG probabilities)
- Step 3: Estimate non-Hispanic white average dealer mark-up
- Step 4: Estimate number of consumers harmed based on number with dealer mark-up greater than non-Hispanic white average
- Step 5: Identify consumers actually harmed

Accuracy issues with BISG proxies

- Lower error rates in some cases than using geography or surname alone
- But still subject to high error rates
  - False positives
  - False negatives
- CFPB's own study found high error rates
- CRA's study found even higher error rates
- Actual degree of error in any given situation is unknown

CFPB's study shows substantial error rates

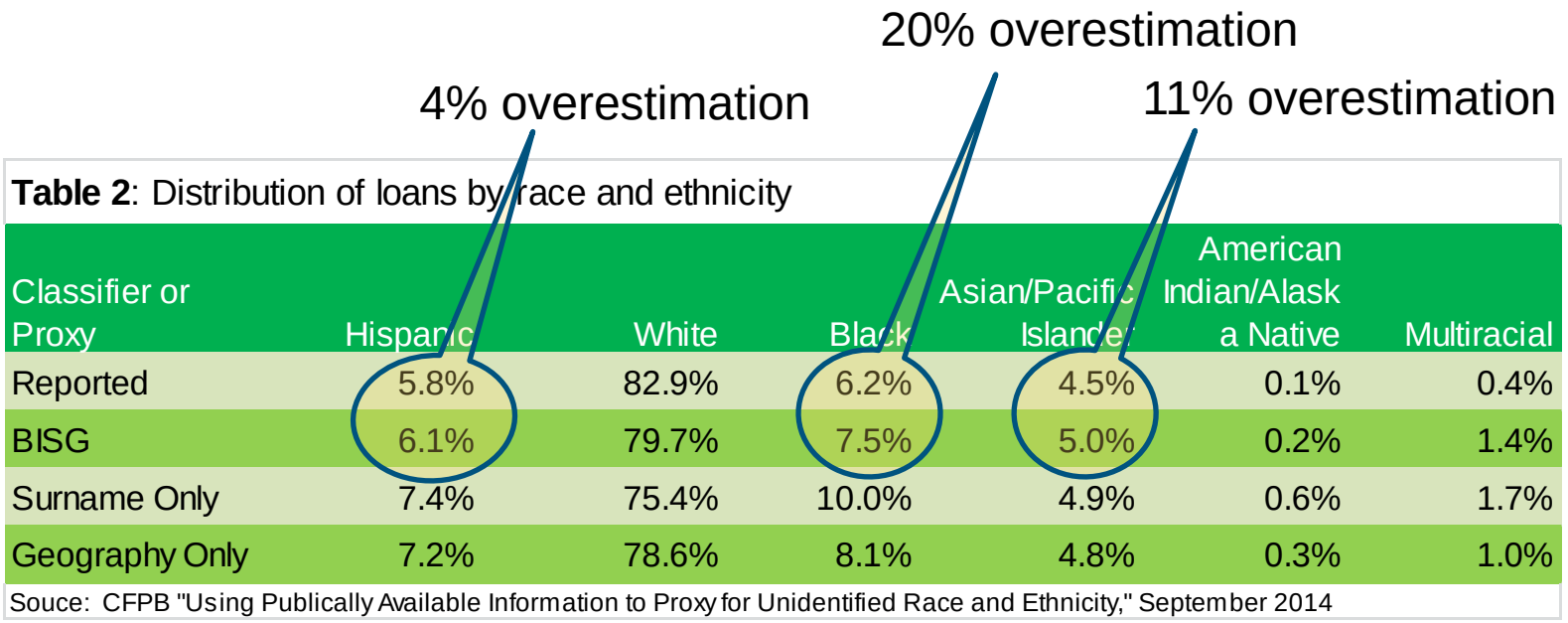


Table 2: Distribution of loans by race and ethnicity

Classifier or Proxy	Hispanic	White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial
Reported	5.8%	82.9%	6.2%	4.5%	0.1%	0.4%
BISG	6.1%	79.7%	7.5%	5.0%	0.2%	1.4%
Surname Only	7.4%	75.4%	10.0%	4.9%	0.6%	1.7%
Geography Only	7.2%	78.6%	8.1%	4.8%	0.3%	1.0%

Souce: CFPB "Using Publically Available Information to Proxy for Unidentified Race and Ethnicity," September 2014

The overestimate of minority populations in CFPB's data set is high under all three methods, but relatively lower with BISG

Based on a sample of 190,435 HMDA mortgage records with surname matches, and proportional race/ethnicity probability measures.

CFPB's study shows substantial error rates

Table 10: Classification Over Ranges of BISG Proxy For Non-Hispanic Black

Black BISG Proxy Probability Range	Total Applications (1)	Estimated Black (BISG) (2)	Reported Black (3)	Reported White (4)	Reported Other Minority (5)
0-10	160,733	1,859	1,466	139,684	19,583
10-20	9,742	1,387	941	8,403	398
20-30	4,916	1,207	906	3,814	196
30-40	3,101	1,072	726	2,242	133
40-50	2,229	997	738	1,408	83
50-60	1,680	922	736	877	67
60-70	1,417	920	765	596	56
70-80	1,407	1,057	963	391	53
80-90	1,517	1,293	1,222	241	54
90-100	3,693	3,548	3,408	200	85
Total	190,435	14,262	11,871	157,856	20,708

Souce: CFPB "Using Publically Available Information to Proxy for Unidentified Race and Ethnicity," September 2014

Many consumers with low probability of being Black are included in estimated count of total Black consumers

Correlations are positive but are they strong enough?

"Correlations associated with the BISG proxy probabilities for Hispanic and non-Hispanic White, Black, and Asian/Pacific Islander are large and suggest strong positive co-movement with reported race and ethnicity. This means, for example, that the Hispanic proxy value is higher on average for individuals who are reported as Hispanic than for those who are not." – CFPB

TABLE 3: CORRELATIONS BETWEEN PROXY PROBABILITY AND REPORTED RACE AND ETHNICITY

Proxy	Hispanic	White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial
BISG	0.81	0.77	0.70	0.83	0.06	0.05
Surname-only	0.78	0.66	0.40	0.81	0.03	0.05
Geography-only	0.45	0.54	0.58	0.38	0.05	0.03

Source: CFPB White Paper, Summer 2014, "Using publically available information to proxy for unidentified race and ethnicity"

CRA's analysis shows larger error rates

Table 9.
Accuracy of Estimate using a Continuous BISG Methodology

Race/Ethnicity	Actual Count	Actual Percent	BISG Count	Average BISG Percent	BISG Error
African American	23,036	7.8%	32,415	11.0%	40.7%
Hispanic	22,004	7.5%	22,200	7.6%	0.9%
Asian	9,662	3.3%	10,028	3.4%	3.8%
Non-Hispanic White	234,746	80.0%	223,031	76.0%	-5.0%

Source: Charles River Associates. HMDA data augmented with proprietary data

Based on 292,000 HMDA records for which self-reported race & ethnicity are known.

Excerpted from "Fair Lending: Implications for the Indirect Auto Finance Market," Prepared by Arthur P. Baines and Dr. Marsha J. Courchane for the American Financial Services Association, Nov. 19, 2014.

Why the error rates and overestimation matter

- If there is a finding of discrimination, BISG inflates the number of affected consumers
  - Result: Inflated damages/restitution claims
- CRA's study showed that BISG proxies together with the CFPB's disparity estimation method can result in substantial upward bias in disparity estimates
  - Results: Possible finding of discrimination where none exists
    Inflated damages/restitution claims

Other issues

- Accuracy declines over time as we get farther from Census measurement years: 2010 for geography demographics and 2000 for surname demographics
- Accuracy can be expected to vary across different types of credit products, and across different economic conditions
- Will become less accurate as society becomes more diverse/integrated
- Does not work with names that are unusual in the US

BISG can introduce bias in disparity estimation

- CRA study found that BISG proxies result in higher disparity estimates compared to using actual self-reported race/ethnicity
- Measurement error tends to be correlated with geography, credit score, and income, among other factors
- Population demographics may not be representative of the demographics of a particular credit product or portfolio
- Standard statistical confidence intervals are inappropriate

Numerous unresolved issues

- There's no single way to do BISG – various options, choices and assumptions, each of which can affect results & accuracy
- Is it sufficiently reliable and tested for establishing liability in enforcement and litigation?
- How to systematically adjust for measurement error & bias?
- How to account for the unknown margin of error in any given application – what's the right "95% confidence interval"?
- How to identify WHO was actually harmed?

Is there any utility in proxy-based fair lending analysis?

- *Must be used given that regulators are using it*
- CFPB justifies the use of BISG proxies based on their positive (albeit far from perfect) correlation with actual race/ethnicity
- Better than nothing??
- Given the correlation, proxy-based analysis can be indicative (directionally) of potential fair lending risk issues if they actually exist
  - Fair lending risk assessment
  - Fair lending monitoring
- Estimates should be interpreted with appropriate caution and humility

References

"Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," Marc N. Elliott, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja & Nicole Lurie, *Health Services Outcomes & Research Methodology*, Vol. 9, 2009, pp. 69–83.

"Using publically available information to proxy for unidentified race and ethnicity," CFPB White Paper, Summer 2014, http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

CFPB statistical software code for its BISG method: https://github.com/cfpb/proxy-methodology

"Fair Lending: Implications for the Indirect Auto Finance Market," Prepared by Arthur P. Baines and Dr. Marsha J. Courchane for the American Financial Services Association, Nov. 19, 2014, http://www.crai.com/sites/default/files/publications/Fair-Lending-Implications-for-the-Indirect-Auto-Finance-Market.pdf

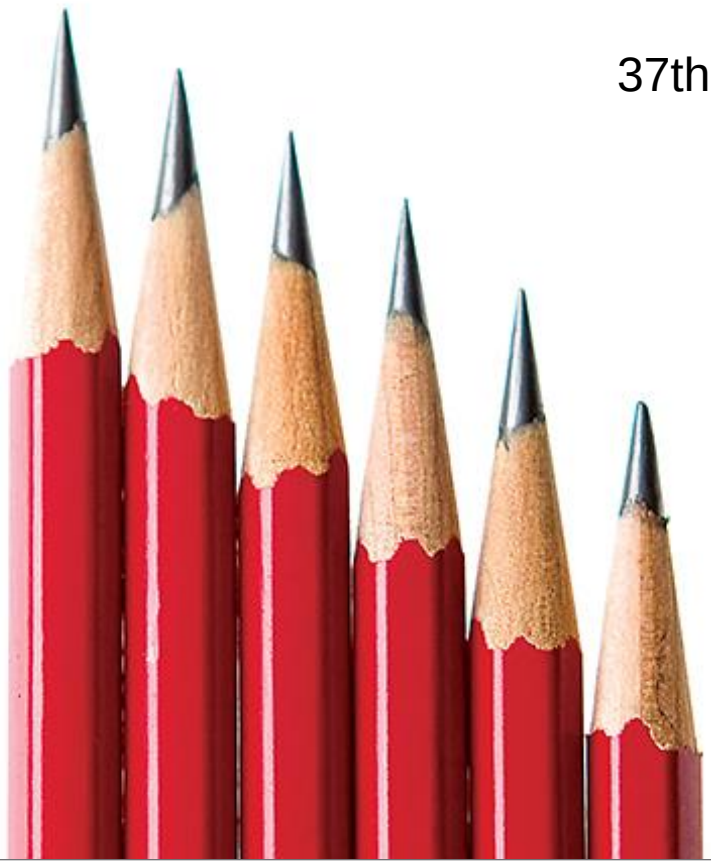
Census Bureau surname list with 2000 Census percentage race/ethnicity frequencies: http://www.census.gov/genealogy/www/data/2000surnames/index.html

"Demographic Aspects of Surnames from Census 2000," David L. Word, Charles D. Coleman, Robert Nunziata and Robert Kominski, U.S. Bureau of the Census, http://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf

Federal Reserve methodology for using Hispanic surname list to proxy ethnicity, and its female and Hispanic names list: http://www.philadelphiafed.org/bank-resources/publications/consumer-compliance-outlook/outlook-live/2013/indirect-auto-lending.cfm

First name/gender data: Social Security Administration, http://www.ssa.gov/oact/babynames/limits.html

Proxy Methods in Fair Lending Analysis



California Bankers Association
37th Annual Regulatory Compliance Conference
October 8, 2015

David Skanderson, Ph.D.
Vice President
dskanderson@crai.com
202.662.3955

CRA Charles River Associates

This presentation is not complete without the accompanying discussion.