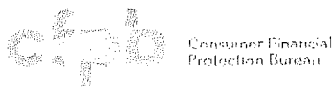


June 4, 2013

Proxy Methodology Discussion

Office of Research



CONFIDENTIAL: SUPERVISORY INFORMATION
SENSITIVE & PRE-DECISIONAL
NOT INTENDED FOR EXTERNAL DISTRIBUTION

Proxies are already used to measure race and ethnicity in a variety of contexts

- Standard practice uses geography for some groups, and surnames for others
 - Each type of information is more/less useful for particular races/ethnicities
- In other applications some businesses use more detailed proxies, incorporating multiple sources of information (e.g., [REDACTED])
 - Methodologies are proprietary and for fee
 - Codes to high granularity with focus on cultural heritage and self-identity
 - Typically used for marketing purposes

CFPB OR proposes a methodology to improve standard practice of proxying for race/ethnicity

- Want to provide a methodology that:
 - relies on sound, foundational statistical principles
 - uses publicly available information
 - improves accuracy of disparity estimates
 - can be refined as “state of the art” improves
 - can be easily implemented
- Proposed methodology consists of unified practice that systematically incorporates the clearest information
- Structure of proxy allows for potential future refinement
 - First name for race/ethnicity
 - Use of Block Groups for geography information

The joint proxy for race and ethnicity proposed by CFPB relies on two sources of information

- Surname
 - US Census list of race/ethnicity by surname for all names with >100 appearances
 - List from 2000 Census published in 2007
- Geography
 - US Census data on race/ethnicity by census tract from 2010 Census
- NOTE: Proxy for gender relies only on Social Security database of infant names by gender
 - Remainder of presentation focuses on joint proxy for race/ethnicity

Bayes' Rule, a well known statistical theorem, generates the joint proxy

- Effectively, the rule combines the knowledge we receive from the two sources of information to refine the probability of an individual belonging to a specific race/ethnicity
 - Informative data (e.g., a name with a high probability of being Hispanic) will heavily impact the combined probability
 - Uninformative data (e.g., living in an area with an equal distribution of races/ethnicities) do not hurt the proxy, only provide little additional refinement

Use of the joint proxy has benefits beyond better estimation of race

Multiple ways exist for thinking about how accurately proxies capture “truth”

1. Correlations between reported race/ethnicity and proxy
2. Distribution of reported race/ethnicity vs. proxy
3. Receiver Operating Characteristics (ROCs)

1. Correlations tell us how much a proxy relates to reported race/ethnicity

- 0=no relationship, 1=perfect co-movement
- Joint proxy is at least as good as (typically better than) alternatives on this metric
- Our results typically match or exceed rates estimated by Elliott

	Joint Proxy	Name Proxy	Geographic Proxy
Hispanic	0.84	0.83	0.49
Asian/Pacific Islander	0.86	0.85	0.47
Black	0.67	0.41	0.53
Non-Hispanic White	0.79	0.72	0.54
Am. Ind/ Native	0.1	0.06	0.07

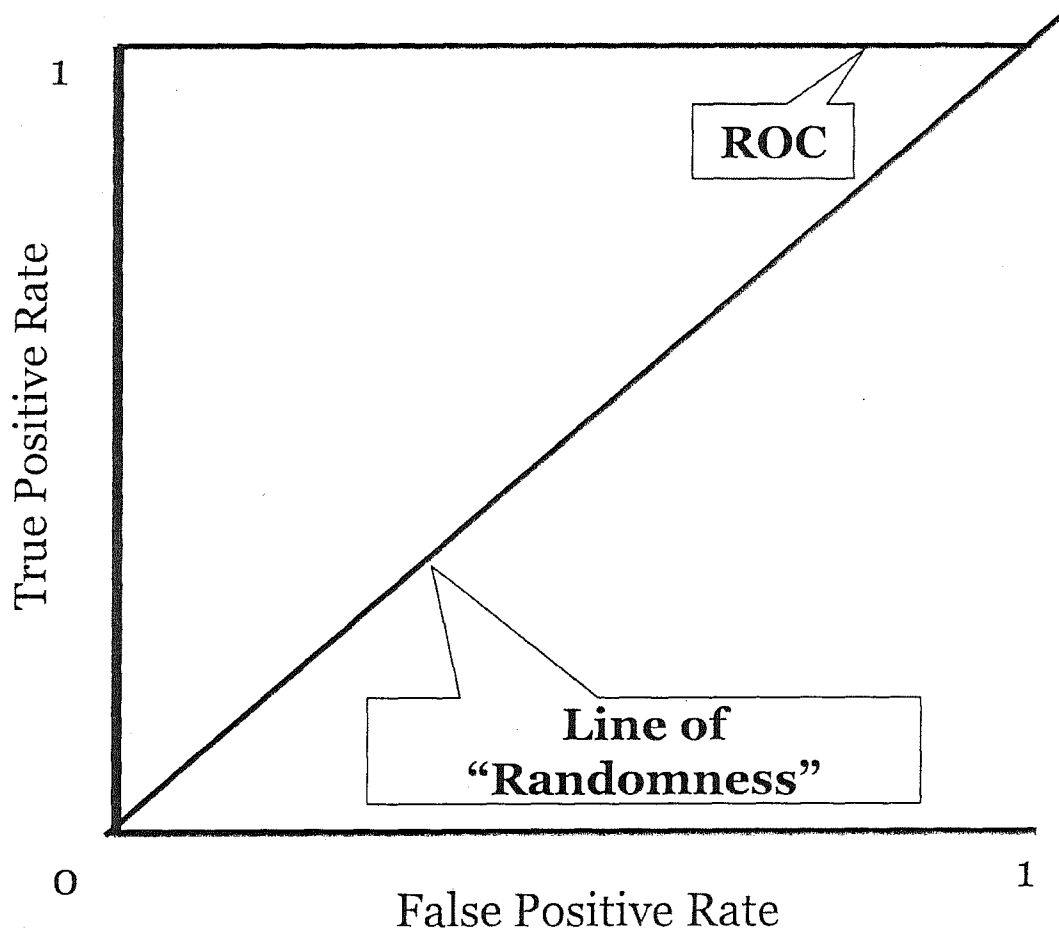
2. The joint proxy more closely matches the distribution of race/ethnicity

	HMDA Reported	Joint Proxy	Name Proxy	Geog. Proxy
Non-Hispanic White	0.73	0.69	0.67	0.67
Black	0.07	0.08	0.1	0.09
Asian/Pacific Islander	0.09	0.09	0.08	0.07
Native	0.003	0.004	0.006	0.005
Hispanic	0.11	0.12	0.12	0.15
Multiracial/Other	0.01	0.02	0.02	0.02

3. ROCs provide both a graphical and statistical measure of accuracy

- Receiver Operating Curves (ROCs) look at how well a proxy is able to sort individuals accurately into a race/ethnicity
- The curve measures the number of false positives and true positives that occur for a given threshold, then graphs those numbers as the threshold moves from 1 to 0 (left to right).
- We can statistically compare different proxies by comparing the areas under each of their ROCs, and see whether they are different

A perfect proxy creates a line that moves along the graph's axes



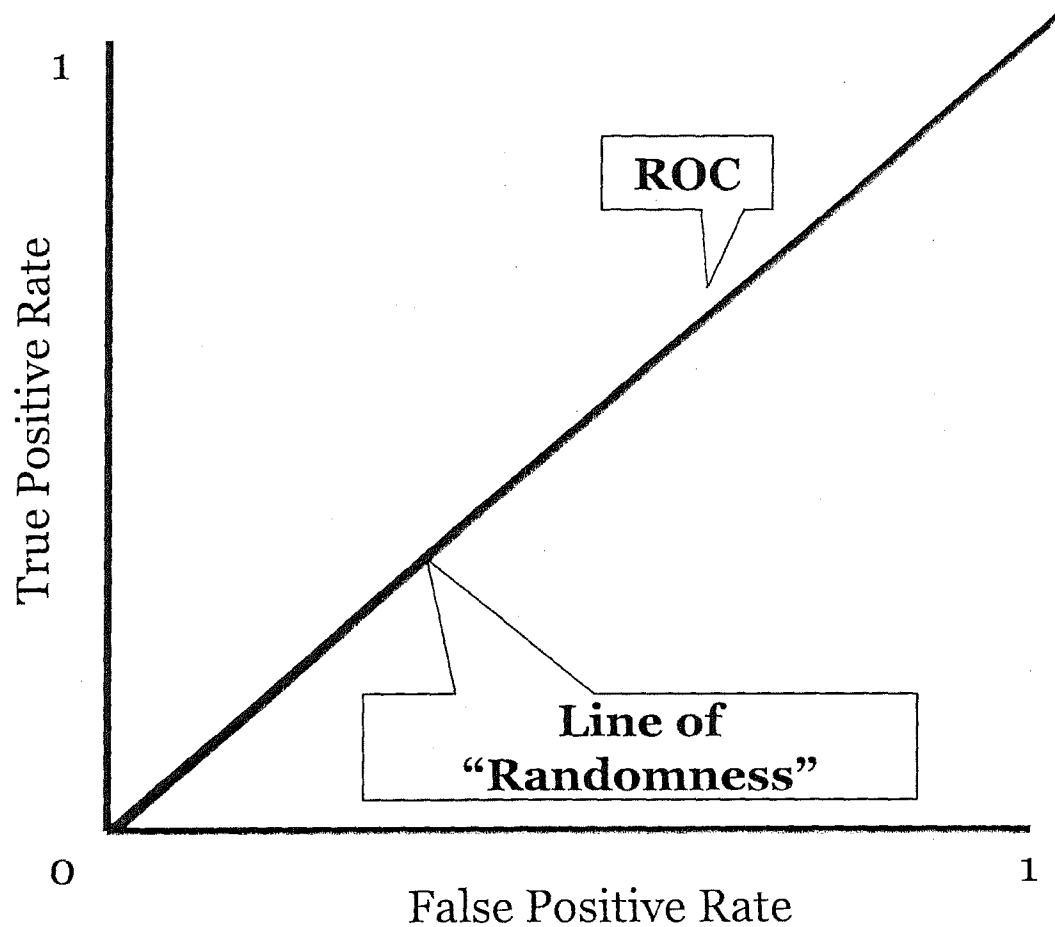
- Y-axis:

$$\frac{\text{True Positives}}{\text{All Actual Positives}}$$

- X-axis:

$$\frac{\text{False Positives}}{\text{All Actual Negatives}}$$

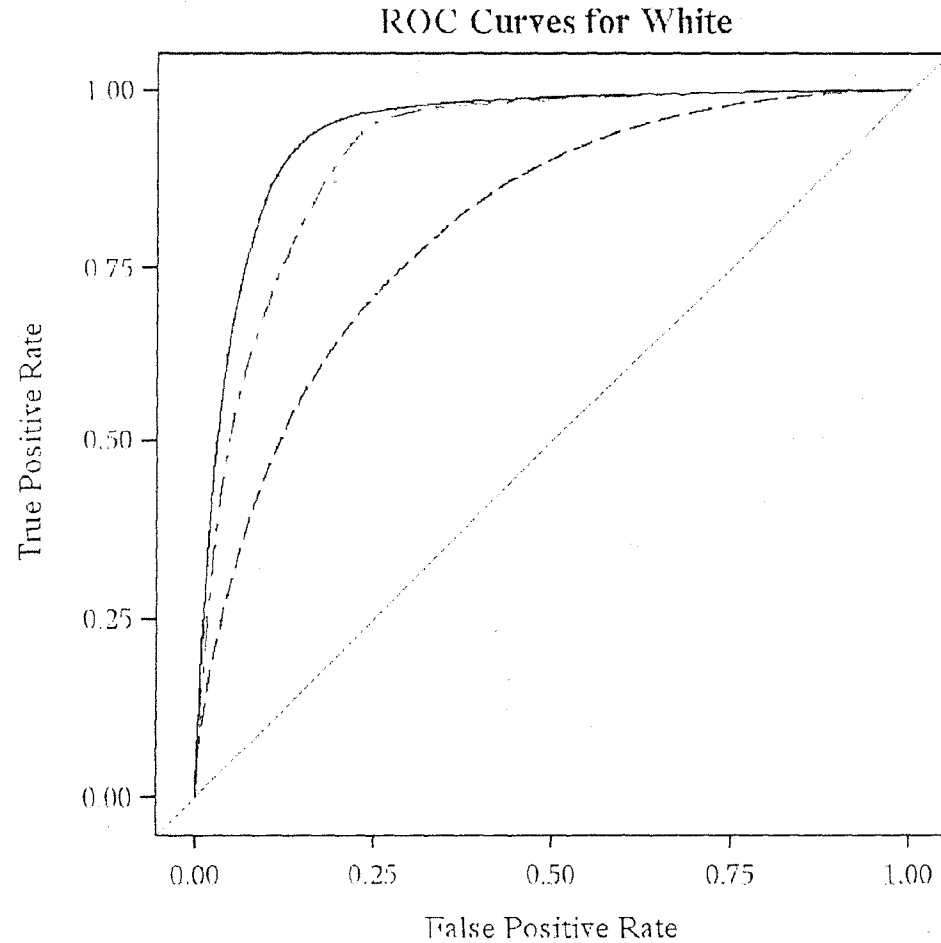
A poor proxy creates a line that moves along the 45-degree line



The joint proxy improves sorting for two groups in particular:

1. Non-Hispanic Whites:

ROC Curve (Area)	
Joint Proxy (0.9430)	Geo Proxy (0.8082)
Name Proxy (0.9154)	



The joint proxy improves sorting for two groups in particular:

2. Blacks:

ROC Curve (Area)	
——— Joint Proxy (0.9442)	----- Geo Proxy (0.8676)
- - - - Name Proxy (0.8657)	

