Using Publicly Available Information to Proxy for Missing Race and Ethnicity: Methodology and Assessment

Contents

1 Introduction

The Equal Credit Opportunity Act (ECOA) and the Consumer Financial Protection Bureau's Regulation B generally "prohibit creditors from requesting and collecting specific personal information about an applicant that has no bearing on the applicant's ability or willingness to repay the credit requested and could be used to discriminate against the applicant," with the notable exception of applications for home mortgages covered under the Home Mortgage Disclosure Act (HMDA).[1] Information on applicant race and ethnicity, however, is often required to conduct fair lending analysis to identify potential discriminatory practices in underwriting and pricing outcomes on the basis of race and national origin.[2]

Comment [BK1]: While the Board has its own Regulation B (for auto dealers), we should refer to the Bureau's Regulation B, which is more generally applicable. Also, the reference in the footnote is to the Bureau's regulation, not the Board's.

Comment [BES2]: Verify that edit is correct.

Comment [BK3]: Let's note that the quotation below is from the Equal Credit Opportunity Act.

Comment [BES4]: Verify citation.

Various techniques exist for addressing this missing data problem. Missing demographic information that reflects group identity—for example, whether or not an individual is White—can be completed by constructing a proxy for the missing information. A proxy may yield a conclusion that a particular individual belongs to a particular group—an individual is classified as being either White or non-White—or may yield group assignment that is probabilistic—an individual is assigned a probability, ranging from 0% to 100%, of being White. When characteristics are not reported for an entire population of individuals, as is usually the case for non-mortgage credit products, techniques focused on completing the missing demographic data generally require relying on additional sources of data and information to construct proxies.

Comment [BES5]: JAL: Not sure "completed" is the right word. Approximated? Estimated?

[1] 12 C.F.R. § 1002.5(a), (b).

[2] The statute makes it unlawful for "any creditor to discriminate against any applicant with respect to any aspect of a credit transaction (1) on the basis of race, color, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract); (2) because all or part of the applicant's income derives from any public assistance program; or (3) because the applicant has in good faith exercised any right under the Consumer Credit Protection Act." 15 U.S.C. § 1691(a).

2 Using census geography and surname data to construct proxies for race and ethnicity

In a variety of settings, including the analysis of administrative health care data and the evaluation of fair lending risk in non-mortgage loan portfolios, researchers and statisticians often rely on publicly available demographic information associated with an individual's surname and place of residence from the U.S. Census Bureau to construct proxies for race and ethnicity when this information is not reported. A proxy for race and ethnicity may be based on the distribution of race and ethnicity within a particular geographic area. Similarly, a proxy for race and ethnicity may be based on the distribution of race and ethnicity across individuals who share the same last name. Traditionally, researchers and statisticians have relied on information associated with either geography or surnames to develop proxies.

Comment [CL6]: Consider introducing here the Fed's methodology as an example of another type of proxy that we are building on.

A recent paper by Elliott et al. (2009) proposes a method to proxy for race and ethnicity that integrates publicly available demographic information associated with surname and the geographic areas in which individuals reside and generates a proxy that is more accurate than those based on surname or geography alone.[3] The method involves constructing a probability of assignment to race and ethnicity based on demographic information associated with surname and then updating this probability using the demographic characteristics of the census block group associated with place of residence. The updating is performed through the application of a Bayesian algorithm, which yields an integrated probability that can be used to proxy for an

[3] Marc N. Elliott et al., Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities, HEALTH SERVICES & OUTCOMES RESEARCH METHODOLOGY (2009) 9:69-83.

individual's race and ethnicity. Elliott et al. (2009) refer to this method as Bayesian Improved Surname Geocoding (BISG).

The Consumer Financial Protection Bureau employs a BISG proxy methodology in its fair lending analysis of non-mortgage credit products and relies on the same public data sources and general methods used in Elliott et al. (2009).[4] The following sections describe these public data sources, explain the construction of the BISG proxy, identify any differences from the general methods used by Elliott et al. (2009), and provide an assessment of the performance of the BISG proxy.

Statistical analysis based on proxies for race and ethnicity is only one factor taken into account by the CFPB in ~~the~~ our fair lending review of non-mortgage credit products. This paper describes the methodology currently employed by the CFPB but does not set forth a requirement for the way proxies should be constructed or used by institutions supervised and regulated by the CFPB. Finally, ~~the~~ our proxy methodology is not static: it will evolve over time as enhancements are identified that improve accuracy and performance.

2.1 Data sources

2.1.1 Surname

[4] The federal banking regulators have made clear that proxy methods may be used in fair lending exams to estimate protected characteristics where direct evidence of the protected characteristic is unavailable. See Interagency Fair Lending Examination Procedures, at 12-13, available at http://www.ffiec.gov/PDF/fairlend.pdf (explaining that "[a] surrogate for a prohibited basis group may be used" in a comparative file review and providing examples of surname proxies for race/ethnicity and first name proxies for sex); CFPB Supervision and Examination Manual, at Procedures 19, available at http://files.consumerfinance.gov/f/201210_cfpb_supervision-and-examination-manual-v2.pdf.

Information used to calculate the probability of belonging to a specific race and ethnicity given an individual's surname is based on data derived from Census 2000 that was released by the U.S. Census Bureau in 2007.[5] This release provides each surname held by at least 100 enumerated individuals, along with a breakdown of the percentage of individuals with that name belonging to each of the six race and ethnicity categories defined by the Office of Management and Budget (OMB): Hispanic; non-Hispanic White; non-Hispanic Black; non-Hispanic Asian/Pacific Islander; non-Hispanic American Indian and Alaska Native; and non-Hispanic Multiracial or Some Other Race.[6,7] In total, the list provides 151,671 surnames, covering approximately 90% of the U.S. population. Word et al. (2008) provides a detailed description of how the census surname list was constructed and describes the routines used to standardize surnames appearing on the list.[8]

Comment [PAF7]: If my assumption is accurate, let's consider dropping a footnote indicating that we will use name tabulations based on the 2010 Census when they are available.

2.1.2 Geography

Information on the racial and ethnic composition of the U.S. population by geography comes from the Summary File 1 ("SF1") from Census 2010, which provides counts of enumerated individuals by race and ethnicity for various geographic area definitions, with census block

[5] The data and documentation are available here: www.census.gov/genealogy/www/data/2000surnames/.
[6] This classification holds Hispanic as mutually exclusive from the race categories, with individuals identified as Hispanic belonging only to that category, regardless of racial background. The Census relies on self-identification of both race and ethnicity when determining race and ethnicity for these individuals, with an exception made for classification to the "Multiracial or Some Other Race" category. In Census 2000, some individuals identifying as "Some Other Race" also specified a Hispanic nationality (e.g., Salvadoran, Puerto Rican); in these instances, the Census identified the respondent as Hispanic.
[7] In the census surname data, the Census Bureau suppressed exact counts for race and ethnicity categories with 2-5 occurrences for a given name. Similarly to Elliott et. al. (2009), in these cases we distribute the sum of the suppressed counts for each surname evenly across all categories with missing nonzero counts.
[8] Word, D.L., Coleman, C.D., Nunziata, R., Kominski, R.: Demographic aspects of surnames from Census 2000. Available at: https://www.census.gov/genealogy/www/data/2000surnames/surnames.pdf.

serving as the finest level of disaggregation.[9] In the decennial Census of the Population, the Census Bureau uses a classification scheme for race and ethnicity that differs slightly from the scheme used by OMB. Census treats Hispanic ~~exists~~ as an ethnicity, ~~while~~ and the other OMB categories ~~are as~~ racial identities. However, Census does report population counts by race and ethnicity in a way that allows for the creation of race and ethnicity population totals consistent with the OMB definition. The CFPB relies on race and ethnicity information for the adult (age 18 and over) population at the census block group, census tract, and 5-digit zip code level, as discussed in the next section.

Comment [PAF8]: This sentence wasn't clear to me so I've suggested edits that hopefully are accurate and helpful.

Comment [BES9]: JAL: Do we want to mention that/how we address this here, so as to not leave the skeptical reader thinking this may be the BISG's Achilles heel?

Comment [BES10]: This is an issue with the way it is written. Follow up with Aaron to determine if counts are reported on SF1.

2.2 Constructing the BISG probability

Constructing the BISG proxy for race and ethnicity for a given set of applicants requires place of residence (address) and name information for those applicants, the census surname list, and census demographic information by census block group, census tract, and 5-digit zip code. The process occurs in a number of steps:

1. Applicants' surnames are standardized and edited, including removing special characters and titles, such as JR and SR, and parsing compound names.
2. Standardized surnames are matched to the census surname list. For applicants with compound surnames, only the first word of the compound surname successfully matched to the surname data is used to calculate the surname based probability. For instance, if an applicant's last name is Smith-Jones, the demographic information associated with Smith

[9] The hierarchy of census geographic entities, from smallest to largest, is: block, block group, tract, county, state, division, region, and nation. Block group level information appears in Table P9 ("Hispanic or Latino, and Not Hispanic or Latino by Race") in the SF1. Table P11 in the SF1 provides similar counts for the restricted population of individuals 18 and over. The public can access these data in a variety of ways, including through the American FactFinder portal at http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml.

is used if Smith appears on the name list. If Smith does not appear on the name list, then the information associated with Jones is used if Jones is on the list.

3. For each name that matches the census surname list, the probability of belonging to a given racial or ethnic group (for each of the six race and ethnicity categories) is constructed. The probability is simply the proportion (or percentage) of individuals who identify as being a member of a given race or ethnicity for a given surname. For example, according to the census surname list, 73% of individuals with the surname Smith report being non-Hispanic White; thus, for any individual with the last name Smith, the surname-based probability of being non-Hispanic White is 73%. For applications with names that do not match the census surname list, a probability is not constructed. These records are excluded in subsequent analysis.[10] Given that approximately 10% of the U.S. population is not included on the census surname list, one would reasonably expect roughly a 10% reduction in the number of records in a proxied dataset due to non-matches to the census surname list.
4. Applicant address information is standardized in preparation for geocoding. Standardization includes basic checks such as removing non-numeric characters from zip codes, making sure zip codes with leading zeroes are accurately identified as such before input into the geocoding algorithm, and ensuring address information is in the correct format.

[10] Elliott et al. (2009) retain records in their assessment data that do not appear on the surname list. To do so, they subtract counts of individuals by race and ethnicity appearing on the name list from the national counts provided in the 2000 Census SF1 and use this distribution to characterize the unlisted population. The CFPB continues to research the approach undertaken by Elliott et al. and may adopt a method for proxying the unlisted surname population in future updates to the proxy methodology.

5. Addresses are mapped into census geographic entities using a geocoding and mapping software application.[11] The geocoding application used by the CFPB identifies the geographic precision to which an address is geocoded, and the precision of geocoding determines the precision of the demographic information relied upon.[12] For addresses that are geocoded to the latitude and longitude of an exact street address (often referred to as a "rooftop"), information on race and ethnicity for the adult population residing in the census block group containing the street address is used; if the census block group has zero population, information for the census tract is used. For addresses that are geocoded to street name, 9-digit zip code, and 5-digit zip code, the race and ethnicity information for the adult population residing in the 5-digit zip code is used. Records with addresses that are geocoded to a precision lower than the 5-digit zip code (for example, city or state) and addresses that cannot be geocoded at all are excluded in subsequent analysis.
6. For geocoded addresses, the proportion (or percentage) of the U.S. adult population for each race and ethnicity residing in the geographic area containing the address or associated with the 5-digit zip code is calculated.
7. Bayes Theorem is used to update the surname-based probabilities constructed in Step 3 with the information on the concentration of the U.S. adult population constructed in Step 6 to create a probability—a value between, or equal to, 0 and 1—of assignment to each of the 6 race and ethnicity categories.

Comment [PAF11]: areas?

Appendix 5 provides the mathematical formula associated with Step 7 and an example of the construction of the BISG proxy probabilities for an individual with the last name Smith residing

[11] The CFPB is currently using ArcGIS Version 10.1 with Street Map Premium 2011 Release 3.
[12] The precision of the geocoding is driven by the availability and the geocoding program's assessment of the quality of address information provided.

in California. The statistical software code, written in Stata, and the publicly available census data files used to build the BISG proxy are available here [insert location and hyperlink]. Because the CFPB currently uses ArcGIS to geocode address information, the geocoding of address information must occur before running the Stata code that builds the BISG proxy. The use of alternative geocoding applications may return slightly different geocoding results and, therefore, may yield different BISG probabilities than those generated using ArcGIS. Finally, Steps 1 through 7 describe the general process currently undertaken by the CFPB to construct proxies for race and ethnicity for fair lending analysis. Unique features of a dataset under review, for example, the quality of surname data and the ability to match individuals to the census surname list or the quality of address information and the ability to geocode to an acceptable level of precision may lead to a modification of the general methodology, as appropriate.

3 Assessing the ability to predict race and ethnicity: an application to mortgage data

Elliott et al. (2009) demonstrate, in the context of a dataset using health plan enrollment data with reported race and ethnicity, that the BISG proxy methodology is more accurate than either the traditional surname-only or geography-only methodologies. In this section, we discuss a similar validation of the BISG proxy in the mortgage lending context.

To assess the performance of the BISG proxy in this context, the geography-only, surname-only, and BISG proxies for race and ethnicity were constructed for applicants appearing in a sample of 190,423 mortgage loan applications in 2011 and 2012 for which address, name, and reported race and ethnicity were provided to the CFPB by a number of lenders pursuant to the CFPB's

Comment [BES12]: The number is

supervisory authority.[13] (We refer to this dataset as the "HMDA" dataset although the dataset includes data elements that are not part of a standard HMDA submission.) Applications with surnames that did not match the surname list and with addresses that could not be geocoded to at least the 5-digit zip code were omitted from the analysis. Table 1 shows that 26,375 observations—approximately 14% of the initial sample (the shaded cells) —were omitted from the analysis, resulting in a final sample of 190,423.

Comment [BES13]: Sample is based on more than one lender.

Table 1 - Mortgage loan sample

		Geocoded	
		No	Yes
Surname match	No	8	26,309
	Yes	58	**190,423**

For each applicant, three probabilities of assignment to each of the six race and ethnicity categories were constructed: a probability based on census race and ethnicity information associated with geography ("geography-only"); a probability based on census race and ethnicity information associated with surname ("surname-only"); and the BISG probability based on census race and ethnicity information associated with surname and geography ("BISG"). As previously discussed, the probabilities themselves may be used to proxy for race and ethnicity by assigning to each record a probability of belonging to a particular racial or ethnic group. These probabilities can be used to estimate the number of individuals by race and ethnicity and to identify potential disparities in outcomes through statistical analysis.

[13] The geography-only probability proxy is constructed in a manner that is similar to the construction of the surname-only proxy. For each geocoded address, the probability of belonging to a given racial or ethnic group (for each of the six race and ethnicity categories) is constructed. The probability is simply the proportion (or percentage) of individuals who identify as being a member of a given race or ethnicity who reside in the block group, tract, or area corresponding to the 5-digit zip code, depending on the precision to which an applicant's address is geocoded.

Assessing the accuracy of the proxy involves comparing a probability that can range between 0 and 1 (a continuous measure) to reported race and ethnicity classifications that, by definition, take on values of only 0 or 1 (a dichotomous measure). Accuracy can be evaluated in at least two ways: (1) by comparing the distribution of race and ethnicity across all applicants based on the proxy to the distribution based on reported characteristics and (2) by assessing how well the proxy is able to sort applicants into the reported race and ethnicity categories. This sorting—the tendency for low values of the proxy to be associated with low incidence of individuals in a particular racial or ethnic group and for high values of the proxy to be associated with high incidence—is measured by the correlation between the proxy and reported classification for a given race and ethnicity. Additional diagnostic measures, such as Area Under the Curve (AUC) statistics, reflect the extent to which a proxy probability accurately sorts individuals into target race and ethnicity and provides a statistical framework for assessing improvements in sorting attributable to the BISG proxy. Section Error! Reference source not found.3.1 provides an evaluation of the use of the BISG probability proxy and assesses performance relative to reported race and ethnicity, illustrating the merits of relying on the BISG probability proxy rather than one based solely on information associated with geography or surname alone.

Comment [BES14]: Consider trying to insert "plain English" descriptions here, though we provide them later.

Comment [BES15]: Provided a bit mor
The AUC is discussed in greater detail in a to... ...ing section.

3.1 Composition of lending by race and ethnicity

Table 2 provides the distribution of reported race and ethnicity ("HMDA reported") and the distributions based on the BISG, surname-only, and geography-only proxies. For the "HMDA reported" row, the percentage in each cell is calculated as the sum of the reported number of individuals in each racial or ethnic group divided by the number of applicants in the sample (multiplied by 100). For the proxies, the percentage is simply the sum of the probabilities for

each race and ethnicity divided by the number of applicants in the sample (multiplied by 100). For example, two individuals each with a 0.5 probability of being Black and a 0.5 probably of being White would contribute a count of 1 to both the Black and the White totals.

Comment [BES16]: JAL: Maybe draw out an example where you can translate the count to a percent, as reported in the table?

Comment [BES17]: Consider tweaking example.

Table 2 - Distribution of loans by race and ethnicity

Classifier or Proxy	Hispanic	Non-Hispanic				
		White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial/ Other
HMDA 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%	75.3%	10.1%	4.9%	0.6%	1.7%
Geography-only	7.2%	78.6%	8.1%	4.8%	0.3%	1.0%

As the chart indicates, all three proxies tend to approximate the HMDA reported population. However, each also tends to underestimate the population of non-Hispanic Whites and overestimate the other race and ethnicity categories, which may reflect differences between the racial and ethnic composition of the census based populations used to construct the proxies and the racial and ethnic composition of individuals applying for mortgages. According to the 2010 Census of Population, 14% of the U.S. adult population was Hispanic; 67% non-Hispanic White; 12% non-Hispanic Black; 5% Asian/Pacific Islander; and 1% American Indian/Alaska Native. According to the 2010 HMDA loan application data for all reporting mortgage originators, only 7% of applications for home mortgages were Hispanic; 80% non-Hispanic White; 6% non-Hispanic Black; 6% Asian/Pacific Islander; and less than 1% American Indian/Alaska Native.[14] Mortgage borrowers tend to be disproportionately non-Hispanic White and, in particular, underrepresent Hispanic and non-Hispanic Blacks relative to the population of the U.S.

Comment [CL18]: What effect does this have on the accuracy of our analyses?

Comment [BES19]: This describes the differences in distributions for the mortgage sample and may not be generalizable to other context. This is addressed, to some degree, in the sentences that follow.

Comment [BES20]: This is for the all ages. Get the 18+ distribution.

Comment [DMS21]: I'm not sure I understand the relevance of these two sentences. Why not go from the first sentence commenting on the results to the next paragraph which elaborates on that comment?

Comment [BES22]: Perhaps put into table format.

Comment [PAF23]: In this and the following paragraph we are making two points that may appear contradictory. On the one hand, we are treating proximity to HMDA reported figures as a measure of accuracy, while on the other hand we are saying that the gap may reflect the fact that the HMDA distribution doesn't match the overall Census distribution. Considering both points could cause one to question whether it's a good thing for Census-based measure to come closer to the HMDA reported measure (e.g., the close % for Hispanics).

I think we can successfully make both points without appearing to contradict ourselves, but probably need to separate them. Here, perhaps we should just focus on the first point, and discuss the fact that the BISG results are closer to HMDA reported that Surname-Only and Geography-Only, and based on that measure we can fairly claim that BISG is more accurate. I've marked a point later in the paper where we can discuss this point, if you agree.

[14] The HMDA distributions for race and ethnicity are based only on applicant information for which race and ethnicity is reported and for applications that were originated, approved but not accepted, and denied by lenders.

Importantly, however, the BISG proxy comes closer to approximating the HMDA-reported population than the traditional proxy methodologies. Though we see small absolute gains in accuracy from use of a BISG proxy for some groups relative to the traditional methods of proxying, these gains frequently represent a sizeable improvement in terms of relative performance. For example, the gap between HMDA reported race and estimated race for non-Hispanic Whites shrinks by 1.1% (from 82.9% – 78.6% = 4.3% to 82.9% – 79.7% = 3.2%) when moving from a geography-only to the BISG proxy. Given the initial gap of 4.3% this represents an almost 25% reduction in the difference between estimated and reported race. The gaps for non-Hispanic Black, non-Hispanic American Indian/Alaska Native, and Hispanic shrink in a similar manner. For non-Hispanic Asian/Pacific Islander, the gap between estimated and reported totals increase by 0.2% in absolute terms compared to the closer geography-only alternative and 0.1% compared to the surname-only alternative. For the non-Hispanic Multiracial/Other category, the BISG proxy does slightly better than the surname-only and slightly worse than the geography-only proxy in approximating the HMDA reported percentage.

3.2 Predicting race and ethnicity for applicants

3.2.1 Correlations between the proxy probability and reported race and ethnicity

Table 3 provides the correlations between reported race and ethnicity and the BISG, surname-only, and geography-only proxies.

Comment [CL24]: Could you provide more information about what a correlation is, either here or above?

Comment [BES25]: We do describe this above and below. I think Chris is asking for a more intuitive explanation.

Table 3 - Correlations between proxy probability and reported race and ethnicity

Proxy	Hispanic	Non-Hispanic: White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial/ Other
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

Correlation is a statistical measure of the relationship between different variables—in this case the race proxy and an individual's actual, reported race. Positive values indicate a positive correlation (as one variable increases in value, so does the other), negative values imply negative correlation (as one variable increases in value, the other decreases), and 0 indicating no statistical relationship. By definition, a correlation coefficient of 0 means that the proxy probability has no predictive power in explaining movement in the reported value, while a coefficient of 1 means that an increase in the proxy probability perfectly predicts increases in the reported values. Higher values of the correlation measure indicate a stronger ability to sort individuals both into and out of a given race and ethnicity classification.

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. For non-Hispanic American Indian/Alaska Native and the Multiracial/Other classifications, correlations are positive but close to zero for all proxy methods, suggesting a low degree of power in predicting reported race and ethnicity for these two groups.

Looking across the rows in Table 3, correlations associated with the BISG are higher than those associated with the surname-only and geography-only proxies, notably for non-Hispanic Black and non-Hispanic White, reflecting the increase in the strength of the relationship between the proxy and reported characteristic from the integration of information associated with surname and geography in the BISG proxy. These results align closely with those found in Elliot et al.

Comment [PAF26]: Can we substitute "accurately predict race or ethnicity" for this phrase to simplify the language? I also note that the opening sentence of the next section (3.2.2) contrasts this analysis with the value of the AUC, which measures the proxy's ability to "successfully sort individuals into each race and ethnicity."

Comment [BK27]: We should consider anticipating the argument that the proxy is only 70% accurate for African Americans.

Comment [BES28]: JAL: Maybe a plain language explanation here? E.g. "This means, for example, that the API proxy value is higher on average for individuals who reports as Asian/Pacific Islanders compared to non-API individuals."

(2009), which, as previously noted, assessed the BISG proxy using national health plan enrollment data.[15]

3.2.2 Area Under the Curve (AUC)

While correlations illustrate the overall extent of co-movement between the proxies and reported race and ethnicity, it is also important to assess the extent to which the proxy probabilities successfully sort individuals into each race and ethnicity. Error! Reference source not found.

A statistic that can be used to calculate this is called the Area Under the Curve (AUC), which represents the likelihood that the proxy will accurately sort individuals into the target race and ethnicity.[16] The AUC has the following interpretation: if one randomly selects an individual who is reported as Hispanic and a second individual who is reported as non-Hispanic, the AUC represents the likelihood that the individual reported as Hispanic has a higher proxy value of being Hispanic than the randomly selected individual reported as non-Hispanic. The AUC can be used to test the hypothesis that one proxy is more accurate than another at sorting individuals in order of likelihood of belonging to a given race and ethnicity. An AUC value of 1 (or 100%) reflects perfect sorting and classification, and a value of 0.5 (or 50%) suggests that the proxy is only as good as a random guess (e.g., a coin toss).

Table 4 provides the results of statistical comparisons of the geography-only, name-only, and BISG probabilities. The AUC statistics associated with the BISG proxy for Hispanic and non-

[15] Table 4 of Elliott et al. (2009): Non-Hispanic White (0.76); Hispanic (0.82); Black (0.70); Asian/Pacific Islander (0.77); American Indian/Alaska Native (0.11); and Multiracial/Other (0.02).

[16] The AUC is based on the Receiver Operating Characteristic (ROC) curve, which plots the tradeoff between the true positive rate and the false positive rate for a given proxy probability over the entire range of possible threshold values that could be used to classify individuals with certainty to the race and ethnicity being proxied. See Appendix X for more detail on the construction of the ROC curves and calculation of the AUC.

Hispanic White, Black, and Asian/Pacific Islander are large and exceed 90%. For instance, the AUC statistic associated with the BISG proxy for non-Hispanic Black is 0.9539, suggesting that a randomly chosen individual reported as Black is 95% more likely to have a higher BISG probability of being Black than a randomly chosen individual reported as non-Black.

Comment [BJK29]: What does this mean? It seems odd to say "95% more likely," since that would imply that a number in excess of 100% is possible. Is it correct to say that the proxy sorts correctly 95% of the time?

Comment [BES30]: JAL: I am a little confused here as well. Could we write something like "an individual who self-reports as Black is almost twice as likely to have a higher BISG value that an individual who reports as non-Black," or "in this case the BISG proxy performs less 5 percentage points worse than perfect sorting"?

Comment [BES31]: Try to build on BJK comment to make more plain English.

Table 4 - Likelihood of assignment of higher proxy probability for group membership given that borrower is reported as member of group (Area Under Curve statistic)

Proxy	Hispanic	Non-Hispanic				
		White	Black	Asian/Pacific Islander	American Indian/Alaska Native	Multiracial/ Other
BISG	0.9447	0.9429	0.9539	0.9723	0.6847	0.6842
Geography-only	0.8387	0.8389	0.8959	0.8359	0.6574	0.6015
Surname-only	0.9303	0.8967	0.8676	0.9651	0.5919	0.7067
p-value, H0: BISG = Geo	<0.0001	<0.0001	<0.0001	<0.0001	0.0219	<0.0001
p-value, H0: BISG = Name	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0318

For each of these four race and ethnicity categories and the non-Hispanic American Indian/Alaska Native category, the AUC for the BISG proxy probability is statistically significantly larger than the AUC for the surname-only and geography-only probabilities, suggesting that, at or above the 99% level of statistical significance, the BISG more accurately sorts individuals than the traditional proxy methodologies. The greatest improvements in the AUC are associated with the BISG proxy for non-Hispanic White and Black, as the AUC is considerably higher than the AUCs associated with the geography-only and surname-only proxies. For Hispanic and non-Hispanic Asian/Pacific Islander, this improvement is only

marginal relative to the performance of the surname-only proxy. Performance for non-Hispanic American Indian/Alaska Native and Multiracial/Other, while generally improved by the use of the BISG proxy probabilities, is weak overall regardless of proxy choice, suggesting that proxies based on census geography and surname data are not particularly powerful in their ability to sort individuals into these two race and ethnicity categories.

Comment [BES32]: JAL: Should we take this opportunity to reiterate how to read these? E.g. "representing only an 18 percentage point (or 36%) improvement in sorting over random chance."

3.2.3 Classification over the range of proxy values

The BISG proxy's ability to sort individuals is made clear through an evaluation of the number of applicants falling within ranges of proxy probability values. For example, for 10% bands of the BISG proxy probability for Hispanics, Table 5 provides: the number of total applicants (column 1); the estimated number of Hispanic applicants based on the summation of the BISG probability (column 2); the number of reported Hispanic applicants (column 3); the number of reported non-Hispanic applicants (column 4); and the number of reported other minority, non-Hispanic applicants (column 5). A few features are worth noting.

Table 5 - Classification over range of BISG proxy for Hispanic

Hispanic BISG proxy probability range	Total Applicants (1)	Estimated Hispanic (BISG) (2)	Reported Hispanic (3)	Reported Non-Hispanic White (4)	Reported Other Minority (5)
0% - 10%	176,093	1,131	1,676	153,953	20,464
10% - 20%	1,727	241	163	1,214	350
20% - 30%	656	163	130	417	109
30% - 40%	541	189	147	312	82
40% - 50%	557	251	226	261	70
50% - 60%	597	328	279	258	60
60% - 70%	803	522	455	264	84
70% - 80%	1,135	853	766	286	83
80% - 90%	1,788	1,529	1,347	347	94
90% - 100%	6,526	6,312	5,883	534	109
Total	190,423	11,519	11,072	157,846	21,505

Estimated Hispanic (BISG) is calculated as the sum of the BISG probabilities for being Hispanic within the corresponding proxy probability range.

First, the distribution of the BISG proxy probability is bimodal with concentrations of total applicants for low (e.g., 0%-10%) and high (e.g., 90%-100%) values of the proxy, which illustrates the sorting feature of the proxy. Reported Hispanic applicants are concentrated within high values of the proxy. For example, 65% ((1,347+5,883)/11,072) of reported Hispanic applicants (column 3) have BISG proxy probabilities greater than 80%; this concentration is mirrored by the estimated number of Hispanic applicants (column 2), for which the same percentage of concentration is 68% ((1,529+6,312)/11,519). While the BISG proxy may assign high values to some non-Hispanic applicants, 97% (153,953/157,846) of the reported non-Hispanic White and 95% (20,464/21,505) of the reported other non-Hispanic minority borrowers have Hispanic BISG proxy probabilities that are less than 10%.

Comment [BES33]: JAL: I found this a little awkward.

Second, over the full range of values of the BISG proxy, there are reported Hispanic applicants; this is also reflected by the estimated counts in column 2. For example, there are 597 applicants with BISG proxy values between 50% and 60%. 279 of these applicants report being Hispanic, while the BISG proxy estimate of the number of Hispanic applicants in this range—calculated again by summing probabilities for individuals within this probability range—is 328.

Comment [DMS34]: Is this correct?

Comment [BES35]: Yes

As suggested by Table 5, the BISG proxy tends to overestimate the number of Hispanic applicants for the mortgage pool under review. In the final row of column (3) we see that the total number of reported Hispanic applicants is 11,072. The estimated total number of Hispanic applicants—calculated as the sum of the BISG probabilities for Hispanic applicants—is 11,519,

which overestimates the number of Hispanic applicants by 4%. This overestimation may reflect, as discussed in Section 3.1, the use of demographic information based on the population at large to proxy the characteristics of mortgage applicants. Recalling that the frequency of Hispanics in the HMDA data is less than half that in the population as a whole, a 4% difference between estimated and reported values represents a relatively small overestimate.

Comment [PAF36]: Here is where I would suggest putting the earlier detail regarding the differences between the HMDA distributions and Census distributions

The CFPB relies directly on the BISG probability in its fair lending related statistical analyses. However in contrast, some practitioners rely on the use of a probability proxy and a threshold rule to classify individuals into race and ethnicity, where individuals with proxy probabilities equal to and greater than a specific value, for example 80%, are considered to belong to a group with certainty, while all others are considered non-members with certainty. Consider two individuals who are assigned BISG probabilities of being non-Hispanic Black: individual A with 82% and individual B with 53%. The application of an 80% threshold rule for assignment would force individual A's probability to 100% and classify that individual as being Black and force individual B's probability to 0 and classify that individual as being White.

Comment [BJK37]: Same comment as above: In whose practice? Are you referring to common industry practice in FL analysis? Or to something else? If the latter, it would be helpful to provide examples and citations.

Comment [BES38]: Rephrased this. A newly added footnote (earlier section) includes reference to the FFIEC guidelines indicating that surrogates can be used. Only the FRB has gone public with their methodology. so could cite to it. Otherwise, familiarity with use in fair lending context is based on what we've observed in our supervisory w

This process The threshold rule removes the uncertainty about group membership at the cost of decreased statistical precision, with that precision deteriorating with decreases in the proxy's ability to create separation across races and ethnicity. In situations in which researchers can obtain clear separation between groups—for instance, situations for which the probabilities of assignment tend to be very close to 0 or 1—the consequences of using a threshold assignment rule, beyond simple measurement error, would be minor. However, when insufficient separation exists—for example, when there are a significant number of individuals with probabilities

between 20% and 80% of belonging to a particular group—the use of thresholds can artificially bias, usually downward, estimates of the number of individuals belonging to particular racial and ethnic groups and potentially attenuate estimates of differences in outcomes between groups. Table 5 makes clear the consequence of applying a threshold rule to the BISG proxy probability to force classification with certainty. If an 80% threshold rule is applied, the estimated number of Hispanic applicants is 8,314—the sum of all applicants in column (1) with a BISG probability equal to or greater than 80%—which underestimates the reported number of 11,072 Hispanic applicants by 25%. The underestimation is driven by the failure to count the large number of individuals in column (3) who are reported as being Hispanic in the mortgage sample but for whom the BISG probability of assignment is less than 80%.

It is worth noting that the application of an 80% threshold rule to classify individuals also yields false positives: individuals who report being non-Hispanic but, nonetheless, are assigned BISG proxy probabilities of being Hispanic equal to or greater than 80%. For the mortgage pool under review, 881 applicants who are reported as being non-Hispanic White and 203 applicants who are reported as being some other minority would be classified as Hispanic by an 80% threshold rule. The false positive rate associated with these 1,084 observations is 0.6%, measured as the number of false positives (1,084) as a percentage of the total number of correctly classified applicants using a threshold rule, which includes the 7,230 true positive reported Hispanics with BISG probabilities greater than or equal to 80% plus 178,267 true negative reported non-Hispanics with BISG probabilities less than 80%. The false discovery rate for these same 1,084 observations is 13%, measured as the number of false positives (1,084) as a percentage of 8,314 applicants identified as Hispanic by the 80% threshold rule.

Comment [BES39]: JAL: Consider moving calculations to footnote (and add formulas). Consider relating this 13% to the overestimation rate.

Classification and misclassification tables for the other five race and ethnicity categories appear in Appendix 6.

4 Conclusion

Information on consumer race and ethnicity is generally not collected for non-mortgage credit products. However, information on consumer race and ethnicity is required to conduct fair lending analysis. Publicly available data characterizing the distribution of the population across race and ethnicity on the basis of geography and surname can be used to develop a proxy for missing race and ethnicity. Historically, practitioners have relied on proxies based on geography or surname only. Recent academic work proposes a new approach—the Bayesian Improved Surname Geocoding (BISG) method—for combining geography- and surname-based information into a single proxy probability. The Consumer Financial Protection Bureau relies on a BISG proxy probability for race and ethnicity in fair lending analysis conducted for non-mortgage products.

Comment [BK40]: We'll probably want to clarify that this is for our supervisory work.

This paper explains the construction of the BISG proxy currently employed by the CFPB and provides an assessment of the performance of the BISG method using a sample of mortgage applicants for which race and ethnicity are reported. Our assessment suggests that the BISG proxy probability is more accurate than a geography-only or surname-only proxy in its ability to predict individual applicants' reported race and ethnicity and generally more accurate than a geography-only or surname-only proxy at approximating the overall reported distribution of race and ethnicity. We also demonstrate that the direct use of the BISG probability does not introduce the sample attrition and significant underestimation of the number of individuals by race and ethnicity that occurs when commonly-relied-upon threshold values are used to classify individuals into race and ethnicity categories.

Comment [BK41]: I presume this is intended, but please note that the prior phrase is "more accurate" and this phrase is "generally more accurate." What explains the difference?

Comment [BES42]: It is intentional.

The CFPB does not require the use or reliance on the specific proxy methodology put forth in this paper, but is making the methodology, statistical software code, and our current understanding of the performance of the methodology for a pool of mortgage applicants available to the public in an effort to foster transparency around our work. Finally, the proxy methodology will evolve over time as enhancements are identified that improve accuracy and performance.

5 Technical appendix – constructing the BISG probability

For race and ethnicity, demographic information associated with surname and place of residence are combined to form a joint probability using the Bayesian updating methodology described in Elliott, et al. (2009). For an individual with surname s who resides in geographic area g:

1. Calculate the probability of belonging to race or ethnicity r (for each of the six race and ethnicity categories) for a given surname s. Call this probability $p(r|s)$.
2. Calculate the proportion of the population of individuals in race or ethnicity r (for each of the six race and ethnicity categories) that lives in geographic area g. Call this proportion $q(g|r)$
3. Apply Bayes' Theorem to calculate the likelihood that an individual with surname s living in geographic area g belongs to race or ethnicity r. This is described by

$$\Pr(r|g,s) = \frac{p(r|s)q(g|r)}{\sum_{r\in R} p * q}$$

Where R refers to the set of six OMB defined race and ethnicity categories. To maintain the statistical validity of the Bayesian updating process, one assumption is required: the probability of residing in a given geography, given one's race, is independent of one's surname. For example, the accuracy of the proxy would be impacted if Blacks with the last name Jones preferred to live in a certain neighborhood more than both Blacks in general and all people with the last name Jones.

Comment [PAF43]: Can we say any more here? Something that affirms that we are comfortable using the methodology even if sociological data indicate that poverty, unemployment, etc. may cause African Americans with the same last name to suffer from reduced mobility and live in the same neighborhoods?

Suppose we want to construct the BISG probabilities on the basis of surname and state of residence for an individual with the last name Smith who resides in California.[17] Table 6

[17] In the example, we choose to use state to make the example more concrete. In practice, a finer level of geographic detail is used as discussed earlier.

provides the distribution across race and ethnicity for individuals in the U.S. with the last name Smith.[18] For individuals with the surname Smith, the probability of being non-Hispanic Black, based on surname alone, is simply the percentage of the Smith population that is non-Hispanic Black: 22.22%.

Table 6 - Distribution of race and ethnicity for individuals in the U.S. population with the surname Smith

Race/Ethnicity	Distribution
Hispanic	1.56%
Non-Hispanic:	
White	73.35%
Black	22.22%
Asian/Pacific Islander	0.40%
American Indian/Alaska Native	0.85%
Multiracial/Other	1.63%

To update the probabilities of assignment to race and ethnicity, the percentage of the U.S. population residing in California by race and ethnicity is calculated. These percentages appear in Table 7.

[18] "Smith" is the most frequently occurring surname in the 2000 Decennial Census of the Population. There are 2,376,206 individuals in the 2000 Decennial Census of Population with the last name "Smith" according to the surname list (www.census.gov/genealogy/www/data/2000surnames/).

Table 7 - Population residing in California as a percentage of the total U.S. population by race and ethnicity

Race/Ethnicity	Population		% of U.S. Population Residing in California
	U.S.	California	
Hispanic	33,346,703	9,257,499	27.76%
Non-Hispanic:			
White	157,444,597	12,461,055	7.91%
Black	27,464,591	1,655,298	6.03%
Asian/Pacific Islander	11,901,269	3,968,506	33.35%
American Indian/Alaska Native	1,609,046	126,421	7.86%
Multiracial/Other	2,797,866	490,137	17.52%
Total	233,564,071	27,958,916	11.97%

Given the information provided in these two tables, we can now construct the probability that Smith's race is non-Hispanic Black, given surname and residence in California using Bayes' Theorem. The probability of being non-Hispanic Black for the surname Smith (22.22%) is multiplied by the percentage of the non-Hispanic Black population residing in California (6.03%) and then divided by the sum of the products of the surname based probabilities and percentage of the population residing in California for all six of the race and ethnicity categories:

$$\frac{.2222 * .0603}{.7335 * .0791 + .0156 * 0.2776 + .2222 * .0603 + .0040 * .3335 + .0085 * .0786 + .0163 * .1605} \approx 16.61\%$$

This same calculation is performed for the remaining race and ethnicity categories. Table 8 provides the surname-only and updated BISG probabilities for all six race and ethnicity categories for individuals with the last name Smith residing in California.

Comment [BK44]: How does this calculation work when the surname suggests a very high likelihood that someone is, say, Hispanic or Asian, but the geography is quite diverse? Is there any way in which Bayes' Theorem allows the 1st piece of information to dominate the estimate, or in other words, stops the 2nd piece of information from clouding an otherwise clear conclusion?

Table 8 - Surname only and BISG probabilities for "Smith" in California

Race/Ethnicity	Surname-only	BISG
Hispanic	1.56%	5.37%
Non-Hispanic:		
White	73.35%	72.00%
Black	22.22%	16.61%
Asian and Pacific Islander	0.40%	1.65%
American Indian/Alaska Native	0.85%	0.83%
Multiracial and Other	1.63%	3.54%

The impact of the adjustment of the surname based probabilities is readily apparent: the surname probability is weighted downward or upward depending on the degree of overrepresentation or underrepresentation of the population of a given race and ethnicity in California relative to the percentage of the U.S. population residing in California. For example, just under 12% of the U.S. population resides in California but nearly 28% of Hispanics in the U.S. reside in California. Knowing that Smith resides in California and that California is more heavily Hispanic than the nation on the whole leads to an increase in the probability that Smith is Hispanic based on surname information alone.

6 Appendix - Additional tables

Table 9 - Classification over ranges of BISG proxy for non-Hispanic White

White BISG proxy probability Range	Total Applicants	Estimated Non-Hispanic White (BISG)	Reported Non-Hispanic White	Reported Minority
	(1)	(2)	(3)	(4)
0% - 10%	20,107	506	2,115	17,992
10% - 20%	3,998	582	940	3,058
20% - 30%	2,745	682	968	1,777
30% - 40%	2,492	871	1,215	1,277
40% - 50%	2,759	1,246	1,605	1,154
50% - 60%	3,348	1,849	2,201	1,147
60% - 70%	4,485	2,930	3,480	1,005
70% - 80%	7,130	5,382	5,874	1,256
80% - 90%	15,665	13,448	14,244	1,421
90% - 100%	127,694	124,289	125,204	2,490
Total	190,423	151,784	157,846	32,577

Table 10 - Classification over ranges of BISG proxy for non-Hispanic African American

African-American BISG proxy probability Range	Total Applicants	Estimated Non-Hispanic Black (BISG)	Reported Non-Hispanic Black	Reported Non-Hispanic White	Reported Other Minority
	(1)	(2)	(3)	(4)	(5)
0% - 10%	160,678	1,863	1,465	139,633	19,580
10% - 20%	9,766	1,391	941	8,426	399
20% - 30%	4,925	1,209	907	3,821	197
30% - 40%	3,104	1,073	726	2,245	133
40% - 50%	2,229	997	737	1,409	83
50% - 60%	1,684	924	736	881	67
60% - 70%	1,419	921	765	598	56
70% - 80%	1,409	1,058	964	392	53
80% - 90%	1,517	1,293	1,222	241	54
90% - 100%	3,692	3,547	3,407	200	85
Total	190,423	14,277	11,870	157,846	20,707

Table 11 - Classification over ranges of BISG proxy for non-Hispanic Asian and Pacific Islander

Asian and Pacific Islander BISG proxy probability Range	Total Applicants	Estimated Non-Hispanic Asian and Pacific Islander (BISG)	Reported Non-Hispanic Asian and Pacific Islander	Reported Non-Hispanic White	Reported Other Minority
	(1)	(2)	(3)	(4)	(5)
0% - 10%	178,490	867	861	154,831	22,798
10% - 20%	1,545	217	235	898	412
20% - 30%	661	161	147	370	144
30% - 40%	495	171	157	250	88
40% - 50%	390	176	145	181	64
50% - 60%	367	202	168	145	54
60% - 70%	415	270	223	160	32
70% - 80%	650	489	421	181	48
80% - 90%	1,268	1,085	923	270	75
90% - 100%	6,142	5,940	5,366	560	216
Total	190,423	9,579	8,646	157,846	23,931

Table 12 - Classification over ranges of BISG proxy for American Indian/Alaska Native

Proxy Probability Range	Total Applicants	Estimated Non-Hispanic American Indian/Alaska Native (BISG)	Reported Non-Hispanic American Indian/Alaska Native	Reported Non-Hispanic White	Reported Other Minority
	(1)	(2)	(3)	(4)	(5)
0% - 10%	190,195	379	238	157,665	32,292
10% - 20%	140	19	3	109	28
20% - 30%	38	9	2	30	6
30% - 40%	12	4	1	9	2
40% - 50%	17	8	1	15	1
50% - 60%	6	3	0	6	0
60% - 70%	5	3	1	4	0
70% - 80%	4	3	1	3	0
80% - 90%	1	1	1	0	0
90% - 100%	5	5	0	5	0
Total	190,423	435	248	157,846	32,329

Table 13 - Classification over ranges of BISG proxy probabilities for Multiracial and Other

Proxy Probability Range	Total Applicants	Estimated Non-Hispanic Multiracial and Other (BISG)	Reported Non-Hispanic Multiracial and Other	Reported Non-Hispanic White	Reported Other Minority
	(1)	(2)	(3)	(4)	(5)
0% - 10%	187,948	2,104	682	156,426	30,840
10% - 20%	1,621	225	34	942	645
20% - 30%	442	107	8	254	180
30% - 40%	198	68	5	114	79
40% - 50%	113	50	9	47	57
50% - 60%	56	31	3	34	19
60% - 70%	33	21	0	18	15
70% - 80%	9	7	0	8	1
80% - 90%	3	2	0	3	0
90% - 100%	190,423	2,615	741	157,846	31,836
Total	187,948	2,104	682	156,426	30,840

7 Technical appendix – Receiver Operating Characteristics (ROC) and Area Under the Curve (AUC)

One way to characterize the proxy's ability to sort individuals into race and ethnicity is to plot the Receiver Operating Characteristic (ROC) curve. The ROC curve is constructed by applying a threshold rule for classification to each race and ethnicity, where probabilities above the threshold yield classification to a given race and ethnicity and those below do not, and then plotting the relationship between the false positive rate and the true positive rate over the range of possible threshold values.

Comment [BK45]: This may be a little confusing, since you've just told the reader that these metrics are more relevant for evaluating threshold proxies, not probability proxies.

Also, it would be helpful to provide more explanation of what exactly the ROC curve means, and in particular the relationship between true positives and false positives.

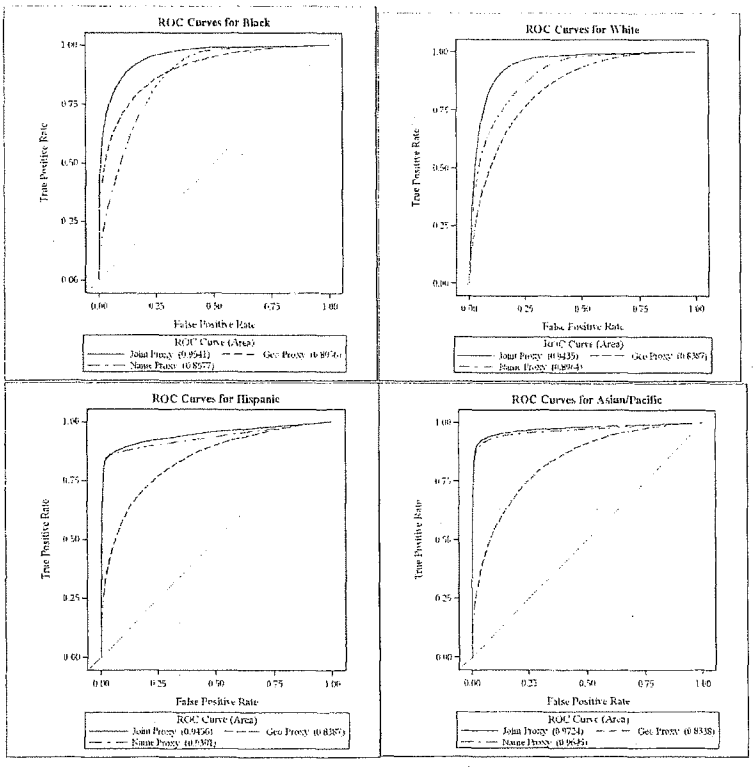
Comment [BES46]: Follow-up with BK

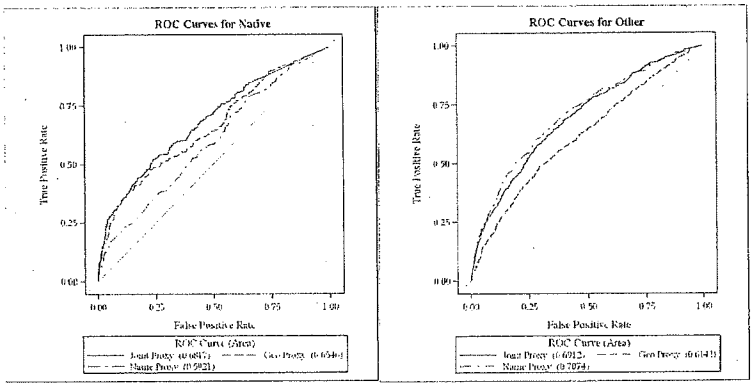
Figure 1 shows the ROC curves for the geography-only, name-only, and BISG probabilities by race and ethnicity. In each plot, the true positive rate is measured on the y-axis and the false positive rate is measured on the x-axis.[19] The slope of the ROC curve represents the tradeoff between identifying the true positives at the expense of increasing false positives over the range of possible threshold values. The ROC curve for a perfect proxy—one that would be able to classify individuals into and out of a given race and ethnicity with no misclassification—moves along the edges of the figure from (0,0) to (0,1) to (1,1). The closer that the ROC curve is to the left and upper edge of the plot area, the better is the proxy at correctly classifying individuals. A proxy that provides no useful information instead moves along the 45-degree line that runs through the middle of the figure. Movement along this line implies that a proxy measure has no ability to meaningfully identify more true members of a group without simultaneously identifying a similar proportion of false positives.

[19] The true positive rate is defined as the ratio of the number of applicants correctly classified into a reported race and ethnicity by a given threshold divided by the total number applicants reporting the race and ethnicity; the false positive rate is defined as the ratio of applicants incorrectly classified into a reported race and ethnicity by a given threshold divided by the total number of applicants not reporting the race and ethnicity.

The graphs demonstrate that for Hispanic and non-Hispanic White, Black, and Asian/Pacific Islander, the BISG proxy is generally associated with a higher ratio of true positives to false positives across all possible threshold values, as shown by the general tendency for BISG's ROC curve to be located to the left and above of the ROC curves for the surname-only and geography-only proxies. The BISG proxy's overall ability to improve sorting, relative to the surname-only or geography-only proxy, is especially notable for non-Hispanic Whites and Blacks. The AUC discussed in Section 3.2.2 is simply the area beneath the ROC curve and above the x-axis.

Figure 1 - Receiver Operating Characteristic (ROC) Curves




ROC Curves for Native
True Positive Rate
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
False Positive Rate
ROC Curve (Area)
ROC Curves for Other
True Positive Rate
False Positive Rate
ROC Curve (Area)