Race, ethnicity, and sex proxy methodology

The indirect auto loan data set provided by [redacted] to the CFPB does not contain information on the race, ethnicity, or sex of applicants. In order to conduct a fair lending review~~wing~~ of pricing and underwriting outcomes, the CFPB developed a proxy methodology for assigning race, ethnicity, and sex to applicants ~~on the basis of~~based on reported address information and name.

Reported addresses for applicants are mapped into census tracts and matched to 2010 Census information on race and ethnicity. Surnames are matched to a list of surnames from the 2000 Census, which reports counts of individuals by race and ethnicity for surnames with at least 100 respondents in that year.[1] Finally, reported first names are matched to a list of first names from the Social Security Administration (SSA), which reports counts of individuals by sex and birth year for first names occurring at least five~~5~~ times in a birth year for a sex~~gender~~.[2]

The demographic distribution of the population across race and ethnicity categories is used to construct a probability (ranging from 0% to 100%) that a given loan application falls into each of the following race and ethnicity categories: non-Hispanic White, non-Hispanic Black, non-Hispanic ~~White,~~ Asian/Pacific Islander, non-Hispanic American Indian/Alaskan Native, and Hispanic.[3] A geography-based probability is constructed on the basis of Census demographic information associated with an applicant's reported address information. A name-based probability is constructed on the basis of Census demographic information associated with an applicant's surname. The demographic distribution of the population across sex categories (male or female) for a given first name is used to construct a probability that a given loan applicant is male or female.

For race and ethnicity, the name and geography information can be combined to form a joint probability using the methodology described in Elliott, et al. [4]~~.~~ This methodology requires the following steps:

- Generate probabilities $p(r|s)$, the probability of belonging to race or ~~/~~ethnicity r given surname s, and $q(g|r)$, the proportion of the population of individuals in race or ~~/~~ethnicity r who live in geographic area g.
- Under the assumption that surname provides no substantial information about where someone lives beyond that provided through race, then through application of Bayes' Theorem and the law of total probability, the likelihood that an individual with surname s living in geographic area g belongs to race or ~~/~~ethnicity r is described by

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

- For applicants with compound surnames or a co-applicant with a different surname, the joint probability that at least one applicant belongs to a given race or ~~/~~ethnicity given all

Comment [BES1]: We may want to strike these assumptions that are implicit in the methodology. These will be highlighted in the technical paper. For the purpose of this description, I think that we just need to describe the procedure.

Formatted: Font:

Formatted: Indent: Left: 0.5", No bullets or numbering

Formatted: Font: Times New Roman, Not Italic

Formatted: Font:

[1] http://www.census.gov/genealogy/www/data/2000surnames/surnames.pdf

[2] http://www.ssa.gov/oact/babynames/limits.html

[3] The construction of race categories that exclude Hispanics (e.g., non-Hispanic Blacks) is driven by the level of reporting in the Census surname list, which separates ethnicity from race.

[4] Elliott et. al., "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities," *Health Services and Outcomes Research Methodology*, Sept. 2009.

surnames ~~was~~ is calculated before incorporating geographic information from the primary applicant. We assume that the ethnicity of one surname does not predict the ethnicity of the other; although this assumption likely does not hold it results in a more conservative estimate of racial identity that underestimates actual ~~disparities~~differences in outcomes. This same method is used for estimating the name-only and ~~gender~~ sex probabilities for joint applicants.

Comment [AS2]: I actually think we should just strike this, but left it in just in case.

Comment [BES3]: My comment regarding the inclusion of assumptions applies here.

Formatted: Font:

Using the geography-based, ~~and~~ surname-based, and joint probabilities for race and ethnicity and the first name-based probabilities for sex, the methodology assigns an application to a given race, ethnicity, and sex by comparing the probability associated with a possible classification to one of three thresholds--70%, 80%, 90%--resulting in up to three possible threshold-based assignments for each demographic classification. For instance, if the probability that an application contains a Hispanic applicant equals 82% then both the 70% and 80% threshold rules would classify the application as Hispanic, but the 90% threshold rule would not.

~~For analytical purposes~~In model estimation, we generate separate results for race, ethnicity, and sex based on ~~all~~ the three threshold rules ~~as well as~~and based on the direct use of the probabilities ~~in model estimation~~.[5] These four general methods—the three threshold methods and the direct use of probability method—are~~were~~ implemented ~~across~~ to explore sensitivities with respect to any estimated ~~disparities~~ differences in outcomes on the basis of race, ethnicity, or~~and~~ sex. When analyzing ~~disparities~~ differences in outcomes using the thresholds, we examine outcomes for individuals assigned to a ~~protected class~~prohibited basis based on a threshold value relative to outcomes for individuals assigned to a control group ~~of individuals~~ based on the same threshold value~~who meet the threshold for likelihood~~. For estimations focused on race and ethnicity, prohibited basis applications are compared to ~~of being~~ non-Hispanic Whites only; ~~in the case of race, and male only for gender~~ for estimations focused on sex, applications with only female applicants are compared to applications with at least one male applicant.[6] ~~To consider disparities for race/ethnicity using the~~When analyzing differences in outcomes across race and ethnicity using the probability values directly, ~~probabi lities directly~~ we ~~incorporate~~ include the ~~likelihood~~ probability value of ~~belonging~~ assignment to each of the prohibited bases ~~protected race/ethnicities~~ in the same regression, with the ~~an~~ implied control ~~of~~ being non-Hispanic White only. For differences in outcomes across sex, we include the probability value of assignment to the female only category, with the implied control being applications with at least one male applicant.

[5] The use of a characteristic assigned on the basis of a threshold rule and the use of a probability directly as a proxy for characteristic are discussed in McCaffrey and Elliott, "Power of Tests for a Dichotomous Independent Variable Measured with Error," *Health Research and Educational Trust,* June 2008.

[6] The probability of being non-Hispanic White application is based on the probability that both the applicant and co-applicant (where applicable) are non-Hispanic White. The probability of being a female only application is based on the probability that both the applicant and co-applicant (where applicable) are female.