**BISG\_RACE\_ETHNICITY**

In conducting fair lending analysis in both supervisory and enforcement contexts, the Bureau’s Office of Research (OR) and Division of Supervision, Enforcement, and Fair Lending (SEFL) rely on a Bayesian Improved Surname Geocoding (BISG) proxy method, which combines geography- and surname-based information into a single proxy probability for race and ethnicity used in fair lending analysis conducted for non-mortgage products. This document describes the steps needed to build the BISG proxies.

The methodology described here is an example of a proxy methodology that OR and SEFL use, although we may alter this methodology in particular analyses, depending on the circumstances involved. In addition, the proxy method may be revised as we become aware of enhancements that would increase accuracy and performance. For more details, see [“Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment”](http://www.consumerfinance.gov/reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/).

Included are a series of Stata scripts and subroutines that prepare the publicly available census geography and surname data and that construct the surname-only, geography-only, and BISG proxies for race and ethnicity. The scripts, subroutines, and data provided here do not contain directly identifiable personal information or other confidential information, such as confidential supervisory information.

Please note that all scripts and subroutines are written for execution in Stata 12 on a Linux platform and may need to be modified for other environments. Users must define a number of parameters, including file paths and arguments for subroutines. The scripts that define the subroutines also identify and describe arguments, as required.

Users must supply their own application- or individual-level data, and any geocoding of those data must occur prior to the execution of the script sequence: **this code assumes that the input application- or individual-level data are already geocoded with census block group, census tract, and 5-digit ZIP code.**

However, included is an example designed to instruct the user in executing the proxy building code sequence. It relies on a set of fictitious data constructed by create\_test\_data.do from the publicly available census surname list and geography data. It is provided to illustrate how the main.do is set up to run the proxy building code and does not reflect any particular individual’s or institution’s information.

A control script, /scripts/main.do, is included to step through the process below. The user will need to change paths and define parameters as required.

1. Geocode the data in a geocoding software package (for example, ArcGIS) to obtain tract and block group identifiers for each record.
2. Build name and geography proxies from Census files included in /input\_files:
   1. Census surname list:
      1. /scripts/surname\_creation\_lower.do—takes .csv file of census surnames, formats surnames to be read as all lower case, and imputes any suppressed values. File created by surname\_creation\_lower.do:
         1. /input\_files/created/census\_surnames\_lower.dta
      2. In order to prepare the user-defined datasets for use with the Census surname list, basic cleaning of surnames using regular expressions and other forms of name standardization is required. This script exists at: /scripts/surname\_parser.do. File created by surname\_parser.do in user-defined directory:
         1. dir'/proxy\_name.dta
   2. Census geographies:
      1. /scripts/create\_attr\_over18\_all\_geo\_entities.do—uses the base information, for individuals age 18 and older, from the Census flat files for block group, tract, and ZIP code[[1]](#footnote-1) and allocates "Some Other Race"[2](https://github.com/cfpb/proxy-methodology#fn-2)[[2]](#footnote-2) to each group in proportion. It creates three files (one each for block group, tract, and ZIP code) with geo probabilities for use in proxy:
         1. /input\_files/created/blkgrp\_attr\_over18.dta
         2. /input\_files/created/tract\_attr\_over18.dta
         3. /input\_files/created/zip\_attr\_over18.dta
3. Calculate the BISG probabilities following the methodology described in [“Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment”](http://www.consumerfinance.gov/reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/).
   1. /scripts/geo\_name\_merger\_all\_entities\_over18.do—this program creates three files (one each for block group, tract, and ZIP code) with BISG probabilities in user-defined directory:
      1. /`maindir'/`inst\_name'\_proxied\_blkgrp.dta
      2. /`maindir'/`inst\_name'\_proxied\_tract.dta
      3. /`maindir'/`inst\_name'\_proxied\_zip.dta
4. The final step is to merge together the block group, tract, and ZIP code-based BISG proxies and choose the most precise proxy given the precision of geocoding, e.g. block group (if available), then tract (if available), or ZIP code (if block group and tract unavailable) using:
   1. /scripts/combine\_probs.do File created by combine\_probs.do in user-defined directory:
      1. /`maindir'/`inst\_name'\_`file'proxied\_final.dta

**Update to proxy methodology – April 2017**

1. In the summer 2014 edition of *Supervisory Highlights*,[[3]](#footnote-3) the Bureau previously reported that examination teams use a Bayesian Improved Surname Geocoding (BISG) proxy methodology for race and ethnicity in their fair lending analysis of non-mortgage credit products. The BISG methodology relies on the distribution of race and ethnicity based on place-of-residence and surname, which are publicly available information from Census. 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 individual’s race and ethnicity.[[4]](#footnote-4)
2. Through March of 2017, examination teams had relied on the surname list derived from the 2000 Decennial Census of the Population in their construction of the BISG proxy for race and ethnicity.[[5]](#footnote-5) In December 2016, the U.S. Census Bureau released a list of the most frequently occurring surnames based on data derived from 2010 Decennial Census of the Population. The updated 2010 list generally uses the same definitions and formats as the list based on the 2000 Census but includes updated values for total counts and race and ethnicity shares associated with each surname.[[6]](#footnote-6) In total, the new surname list provides information on the 162,253 surnames that appear at least 100 times in the 2010 Census, covering approximately 90% of the population.[[7]](#footnote-7) While 146,516 names appear on both the 2000 and 2010 surname lists, the 2010 list contains 15,737 names that do not appear on the 2000 list, whereas the 2000 list contains 5,155 names that do not appear on the 2010 list.[[8]](#footnote-8)
3. As of April 2017, examination teams are relying on an updated proxy methodology that reflects the newly available surname data from the Census Bureau. Our updated proxy methodology relies on the race and ethnicity shares for the 162,253 names that appear on the 2010 list and supplements this list with the race and ethnicity shares for the 5,155 names that appear on the 2000 list but not on the 2010 list, resulting in a list of 167,409 surnames in total.[[9]](#footnote-9)

The updated name list; statistical software code, written in Stata; and other publicly available data used to build the BISG proxy are available at <https://github.com/cfpb/proxy-methodology>.

Please direct all questions, comments, and suggestions to: [CFPB\_proxy\_methodology\_comments@cfpb.gov](mailto:CFPB_proxy_methodology_comments@cfpb.gov).

1. When referring to ZIP code demographics, we match the institution-based ZIP code information to ZIP Code Tabulation Areas (ZCTAs) as defined by the U.S. Census Bureau. [↑](#footnote-ref-1)
2. In the 2010 SF1, the U.S. Census Bureau produced tabulations that report counts of Hispanics and non-Hispanics by race. These tabulations include a “Some Other Race” category. We reallocate the “Some Other Race” counts to each of the remaining six race and ethnicity categories using an Iterative Proportional Fitting procedure to make geography based demographic categories consistent with those on the census surname list. [↑](#footnote-ref-2)
3. *See* Consumer Financial Protection Bureau, *Supervisory Highlights: Summer 2014* (Sept. 17, 2014), *available at* http://files.consumerfinance.gov/f/201409\_cfpb\_supervisory-highlights\_auto-lending\_summer-2014.pdf. [↑](#footnote-ref-3)
4. For more information on the methodology, *see* Consumer Financial Protection Bureau, *Using publicly available information to proxy for unidentified race and ethnicity* (Sept. 2014), <http://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf>. [↑](#footnote-ref-4)
5. *See id.* [↑](#footnote-ref-5)
6. For more details on the updated 2010 surname list, including revisions to the 2000 methodology and programming, *see* Joshua Comenetz, *Frequently Occurring Surnames in the 2010 Census* (Oct. 2016), <http://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf>. [↑](#footnote-ref-6)
7. The surname data are available on the Census Bureau’s website, *see* Frequently Occurring Surnames from the 2010 Census (last revised Dec. 27, 2016), <https://www.census.gov/topics/population/genealogy/data/2010_surnames.html> [↑](#footnote-ref-7)
8. Names must appear at least 100 times in the 2010 Decennial Census in order to be included on the surname list. [↑](#footnote-ref-8)
9. Although these names are not on the 2010 list, and thus likely no longer meet the 100-name threshold, we chose to include them so as to incorporate as much available surname information as possible into the proxy. [↑](#footnote-ref-9)