READ ME FOR /BISG\_RACE\_ETHNICITY

This document describes the steps needed to build the Bayesian Improved Surname Geocoding Method (BISG) proxies.

The Consumer Financial Protection Bureau (CFPB) employs a BISG proxy methodology in its fair lending analysis of non-mortgage credit products 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 proxy methodology is not static: it will evolve over time as enhancements are identified that improve accuracy and performance. See “Using Publicly Available Information to Proxy for Missing Race and Ethnicity: Methodology and Assessment,” for more details.

Included are a series of Stata scripts and subroutines that prepare the input public use 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 do not contain Personally Identifiable Information (PII) or Confidential Supervisory Information (CSI).

Please note that all scripts and subroutines are written in STATA 12 on Linux 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.

A control script, /BISG\_RACE\_ETHNICITY/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 /BISG\_RACE\_ETHNICITY/input\_files:
   1. Census surname list:
      1. /BISG\_RACE\_ETHNICITY/scripts/surname\_creation\_lower.do—takes .csv file of census surnames, formats surnames to be read as all lower case, and reallocates suppressed percentages. File created by surname\_creation\_lower.do:
         1. /BISG\_RACE\_ETHNICITY/input\_files/created/census\_surnames\_lower.dta
      2. In order to prepare the user-defined datasets for use with the Census surname list, it is required to do basic cleaning of surnames using regular expressions and other forms of name standardization. This script exists at: /BISG\_RACE\_ETHNICITY/scripts/surname\_parser.do

File created by surname\_parser.do in user-defined directory:

* + - 1. `dir'/proxy\_name.dta
  1. Census geographies:
     1. /BISG\_RACE\_ETHNICITY/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 and allocates "Some Other Race" 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. /BISG\_RACE\_ETHNICITY/input\_files/created/blkgrp\_attr\_over18.dta
        2. /BISG\_RACE\_ETHNICITY/input\_files/created/tract\_attr\_over18.dta
        3. /BISG\_RACE\_ETHNICITY/input\_files/created/zip\_attr\_over18.dta

1. Calculate the BISG probabilities following the methodology described in “Using Publicly Available Information to Proxy for Missing Race and Ethnicity: Methodology and Assessment” using the probabilities in the previously calculated (above) name and geography files:
   1. /BISG\_RACE\_ETHNICITY/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"
2. 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. /BISG\_RACE\_ETHNICITY/scripts /combine\_probs.do

File created by combine\_probs.do in user-defined directory:

* + 1. /"`maindir'/`inst\_name'\_`file'proxied\_final.dta"