Technical Documentation

Oregon Criminal Justice Commission

Probabilistic Race Correction for Hispanics

September 2018

Several data sets that the Oregon Criminal Justice Commission (CJC) relies upon have known discrepancies between third-party reported race/ethnicity values and self-reported race/ethnicity values. In particular, third-party reported race variables include a higher proportion of White observations and lower proportions of Hispanic, Native American, and Asian when compared to self-reported information. This appendix describes the quantitative method that CJC employs to reduce these discrepancies.

Summary of the problem

A sample of 5000 individuals in the Department of Corrections (DOC) system were surveyed in 2015. The survey included a question about the individual's race/ethnicity, resulting in the DOC dataset containing both a third-party reported race variable and the individual's self-reported race for this sample. Table 1 compares the observed and self-reported variables and provides a snapshot of the erroneous race assignment problem. In particular, observed race overestimates White by 773 observations or about 15% of the total survey sample when compared to the self-report information. Of the 773 erroneously assigned White observations, 320 self-identify into the Hispanic category and 196 identify as Native American categories (see row 1, "White", of Table 1). Conversely, only 12 observations of the 3100 self-reported White observations are misallocated to other categories.

The scale of this measurement error is significant at the full population level. At the time this document was created, the DOC data, for example, has approximately 442,000 unique id numbers, about 90 times the size of the surveyed sample. If the same proportions are found in the broader population, this suggests that roughly 66,000 observations are improperly included in the White category. Further, these discrepancies in the race variable are not limited to the DOC data set but are present in all data sets that do not rely on self-reported race.

Table 1 – Observed and self-reported race for 2015 survey sample

	Self-reported race							
Observed Race	Total	White	Black	Hispanic	Asian	Native	Other	Unknown/ No Answer
Total	4,904	3,100	340	758	97	289	159	161
White	3,873	3,088	21	320	28	196	112	108
Black	430	3	317	17	5	15	42	31
Hispanic	430	4	0	406	3	2	0	15
Asian	63	1	1	1	57	1	2	0
Native	102	4	1	12	3	75	3	4
Unknown	6	0	0	2	1	0	0	3

Probabilistic Matching Methods

One approach to reduce the measurement error in the race variable is by using publically available demographic information associated with surnames, first names, and the geographic areas of residence. (Elliott et al. 2009) first proposed a general method of relying on surnames and race at the census block group level, which is referred to as Bayesian Improved Surname Geocoding (BISG). Since its introduction, the BISG approach has been applied to the study of administrative health care data, in studies and litigation evaluating mortgage and non-mortgage lending patterns, in academic research, and by financial institutions. Recent efforts (Tzioumis 2018; Voicu 2018) have expanded this method to also include first name demographic information based on a mortgage application data set, dubbing this method Bayesian Improved First name Surname Geocoding (BIFSG).

The BIFSG approach utilizes three data sources for matching with the names and location of processing of individuals found within CJC data sets:

US Census 2010 Surname Database

Following both the 2000 and 2010 Decennial Censuses, the US Census Bureau compiled a database of surnames broken down by racial identity. Based on data derived from the Census more generally, this database reports the share of individuals identifying with a given racial category for all surnames with at least 100 enumerated individuals. This accounts for over 294,979,229 individuals across the United States, or nearly 96 percent of the US population. The possible racial/ethnicity categories include non-Hispanic White, Hispanic, non-Hispanic African American, non-Hispanic Asian or Pacific Islander, non-Hispanic American Indian/Alaska Native, and non-Hispanic multirace. For the purposes of this project, these categories are collapsed to White, African American (Black), Hispanic, Asian, Native, and Other categories. For technical documentation regarding the construction of this database, please refer to Comenetz (2016).

US Census Geographic Data

Geographic data regarding the racial and ethnic composition of the U.S. population by race originates in the Census Summary File 1 (SF1). The SF1 file can be used to calculate the racial distribution of a variety of geographic areas, including blocks, block groups, tracts, counties, states, regions, and the nation. While utilizing data from the highest level of disaggregation is preferred, data limitations required the CJC utilize racial distribution data aggregated at the county level.

Mortgage Application First Name Data

Tzioumis (2018) compiles information from three proprietary mortgage datasets to develop probabilities of each race/ethnicity category for each of 4,250 first names that occur at least 30 times in the data. Probabilities are developed for each of the same six race/ethnicity categories found in the Census data. In total, the probabilities are based on a total of 2,663,364 first name observations. For technical documentation regarding the construction of this database, please refer to Tzioumis (2018)

The BIFSG approach utilized in this briefing was applied using the following steps:

1. **Surname Standardization**. Special characters, suffixes, titles, and hyphens were removed, and compound names were parsed.

2. **Surname Matching**. Utilizing the 2010 US Census Surname database, researchers matched the records from Oregon DOC. For compound names, both names were matched where possible.

3. **Surname Race Probabilities**. Where a surname match occurred, the probability of belonging to a given racial group was constructed. This probability was merely the proportion of individuals across the United States with the last name in question who identified as members of a given race. In the event a last name is not matched to the Census Surname list, no probabilities are assigned for first name.

4. **First Name Standardization**. Special characters, suffixes, titles, and hyphens were removed, and compound names were parsed.

5. **First Name Matching**. Utilizing the Mortgage Application database, researchers matched the records from Oregon DOC. For compound names, both names were matched where possible.

6. **First Name Race Posterior Probabilities.** Where a first name match occurred, the probability of the first name within each race group was derived. Bayesian updating of the base surname probabilities requires the calculation of posterior probabilities for each of first name and geography. The probability of a given race for a surname-first combination is the probability of that race conditional on surname multiplied by probability of the first name conditional on the given race. In the event that a first name is not matched to the Mortgage Application database, no probabilities are assigned for first name.

7. **Geographic Race Proportions and Matching**. Using the Census SF1 file, racial distributions were created for each of the 36 counties in Oregon. Similar to Step 6, posterior probabilities are calculated for each county-race combination.

8. **Construction of BIFSG Probabilities**. Bayes theorem was used to update the surname-based probabilities created in Step 3 using the first name probabilities from Step 6 and the geographic information in Step 7. This technique took the following form:

$$\Pr(r|s,f,g) = \frac{p_s(r|s) \times p_f(f|r) \times p_g(g|r)}{\sum_{r \in R}\left(p_s(r|s) \times p_f(f|r) \times p_g(g|r)\right)}$$

$$where\ R = White, Black, Hispanic, Asian, Native, Other$$

where $p_s(r|s)$ represents the probability of belonging to a given race/ethnicity for a given surname, $p_f(f|r)$ represents the probability of having a first name given the racial group, $p_g(g|r)$ represents the probability of being in a county given the racial group, and *R* represents the set of six race/ethnicity categories. The result is a probability for each race category based on surname, first name, and geographic concentrations for each of the six race and ethnicity categories for each observation.

9. **Adjusting Probabilities to County Baseline.** Oregon has a relatively high White population and relatively low non-White populations, which is not mirrored in the DOC population. As a result, iterations of the BIFSG approach revealed that significant misallocations were occurring. A significant performance improvement is made by assigning the race that has the highest probability relative to the county probability from Step 7. This further adjustment leads to a roughly 1% additional reduction in the overall error rate.

10. **Updating DOC Data**. BIFSG and similar methods are often applied to datasets with no information regarding race. Data collected by the Oregon DOC, however, provides information regarding the third-party identification of individual inmates' race. The BIFSG approach can, thus, be used to augment the DOC data rather than as the primary method for assigning racial identity. As such, the following reassignment steps were taken[1]:

    a. Individuals identified as Hispanic based on their BIFSG probability adjusted to the county baseline and with third-party designation of White are assigned as Hispanic.

    b. Individuals identified as Unknown in DOC data are reassigned to the highest probability race category relative to the county baseline, if highest probability category is White or Hispanic.

Table 2 – Observed and self-reported race for 2015 survey sample

	Self-reported race							
Observed Race	Total	White	Black	Hispanic	Asian	Native	Other	Unknown/ No Answer
Total	4,904	3,100	340	758	97	289	159	161
White	3,873	***3,088***	21	***320***	28	196	112	108
Black	430	3	317	17	5	15	42	31
Hispanic	430	***4***	0	***406***	3	2	0	15
Asian	63	1	1	1	57	1	2	0
Native	102	4	1	12	3	75	3	4
Unknown	6	0	0	2	1	0	0	3

Table 2 – BIFSG-corrected race and self-reported race for 2015 survey sample

	Self-reported race							
BIFSG Race	Total	White	Black	Hispanic	Asian	Native	Other	Unknown/ No Answer
Total	4,904	3,100	340	758	97	289	159	161
White	3,579	***3,046***	20	***109***	28	185	105	86
Black	430	3	317	17	5	15	42	31
Hispanic	729	***46***	1	***619***	3	13	7	40
Asian	63	1	1	1	57	1	2	0
Native	103	4	1	12	3	75	3	4
Unknown	1	0	0	0	1	0	0	0

[1] A discussion of these steps can be found in the following section.

Discussion

Table 2 presents the results of the BIFSG correction on the 2015 survey sample. Table 1 is presented again for comparison purposes. The overall error rate across all race categories declines by 3.43%, from 19.54% to 16.11%, which is almost entirely due to the movement from the White category to Hispanic, the most problematic group in the sample data.

- Misidentification of Hispanic individuals as White drops by 211, from 320 ⇒ 109
- Misidentification of White individuals as Hispanic increases by 42, from 4 ⇒ 46
- Correct identification of Hispanic increases by 213, from 406 ⇒ 619 (52.46% improvement)
- Correct identification of White decreases by 42, from 3,088 ⇒ 3,046 (1.36% decline)

Corrections are only applied to the Hispanic category because of BIFSG's poor performance in other categories. Several adjustments to Step 10 of the algorithm were compared to the performance of the BIFSG approach with adjustments only to the White and Hispanic categories, but this strongest performing allocation rule.

A general pattern emerged when comparing these algorithms: predicted population distributions (the Total column) almost always grew closer to the true population proportions, but nearly all of the improvements to the aggregated totals occurred with misidentified individuals. A hypothetical example: Total Native population increases from 103 to 150, but the 47 additions all self-identify as White or Black. The exception to this erroneous pattern among the algorithms is the reassignment to the Hispanic category, where correct reidentification swamps incorrect reidentification by a ratio of more than 4:1 with the BIFSG algorithm and the total count of Hispanic individuals remains lower than in the self-identified sample population.

Additional extensions have also been considered. BISG algorithms were compared to BIFSG, but yielded neither an equivalent reduction in the overall error rate nor equivalent improvement in true positive Hispanic identification. Similarly, a BIFS approach that omitted geographic information was also explored, but yielded a similarly relatively poor performance when compared to BIFSG. In all cases the algorithms were evaluated both using the Step 9 county baseline adjustments and omitting these adjustments in two different iterations. The BIFSG with step 9 included remained the best performing approach through all these comparisons.

Future Extensions

Finally, a fuzzy matching algorithm was developed that finds the closest matching name for both surnames and first names, respectively, in cases where an exact match is not present. The fuzzily matched name is then assigned probabilities (Steps 2-6) where there were no probabilities for the original name. Fuzzy matching in this way can be helpful where there may be errors in user entry or when a name is spelled with an uncommon variant from the original spelling. These methods do yield further efficiency gains, but are computationally intensive and can take extended periods of time with large datasets. For the current survey sample of about 5000 observations the number of individuals falsely identified as Hispanic decreased by 6, positive identifications as Hispanic decreased by 4, and the error rate across all races decreased by a further 1%. While there are some improvements introduced with the fuzzy matching algorithm, the algorithm also introduces additional error and can take close to a day to complete with some of the larger data sets available to the CJC.

Bibliography

Comenetz, Joshua. 2016. “Frequently Occuring Surnames in the 2010 Census.” United States Census Bureau.

Elliott, Marc N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. “Using the Census Bureau’s Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities.” *Health Services and Outcomes Research Methodology* 9 (2): 69–83. https://doi.org/10.1007/s10742-009-0047-1.

Tzioumis, Konstantinos. 2018. “Demographic Aspects of First Names.” *Scientific Data* 5 (March): 180025. https://doi.org/10.1038/sdata.2018.25.

Voicu, Ioan. 2018. “Using First Name Information to Improve Race and Ethnicity Classification.” *Statistics and Public Policy* 5 (1): 1–13. https://doi.org/10.1080/2330443X.2018.1427012.