A Short Analysis on Male Matched Databases

Xiaoyun QIU

6/25/2017

How do the male matched datasets look like? Are all blocked files matched? If not, Why some of them cannot be matched? Do the matched rate differ among different race, gender, birth place combination? If so, what are the characteristics? This short analysis aims to answer these questions.

1. Blocked Files Not Mathced

1.1 20-30 Male

The biggest distinction between 1920 and 1930 data is that the race category mulatto is cancelled in 1930. Therefore, many mulatto files in 1920 are not matched in the 20-30 combination. And some of these files are big. The biggest file, Virginia mulatto male, has 85,987 observations, while in total, there are 747,810 mulatto males in 1920 census data. For other race and birth place combinations in 1920 that are not matched, most of them are small files with less than 1000 observations. Please refer to "/disk/bulkw/sun.lee/matches/matches2030-1920notin1930.csv" for detailed information. For 1930 race and birth place combinations that are not matched in 20-30 combinations, meaning that the race-birth place combinations are not found in 1920, most of them are files with small sizes. Please refer to "/disk/bulkw/sun.lee/matches/matches2030-1930notin1920.csv" to check 1930 race-BPL combinations that are not in 1920.

However, among these not matched race-BPL combiantions, some of them are showing in only one year while some of them are showing in both years. For combinations showing in only one year, in theory, they could not be matched. For those combinations showing in both years, I want to know how they look like. If they are small files, then one explanation of why they cannot be matched is that due to mortality, name changed and transcription errors, those in the race-BPL category disappear in the latter census records. "/disk/bulkw/sun.lee/matches/matches2030-common2030_notmatched2030.csv" contain all the common race-BPL categories that are not matched after pruning. Either the race-BPL categories are small in one of the year or they are small in both years, which is consistent with my understanding.

1.2 30-40 Male

Race categories are quite consistent in 1930 and 1940. The race-BPL combinations not matched in 1930 and 1940 are small files. Please refer to "/disk/bulkw/sun.lee/matches/matches3040-1930notin1940.csv" to check race-BPL combinations showing in 1930 but not showing in 1940, and "/disk/bulkw/sun.lee/matches/matches3040-1940notin1930.csv" to check race-BPL combinations showing in 1940 but not showing in 1930.

As for common categories in 1930 and 1940 that are not matched, similar conclusion applies here. Refer to "/disk/bulkw/sun.lee/matches/matches3040-common3040_notmatched3040.csv". For race-BPL categories that are relatively lager in one year than another, one thing should be paid attention is that the classification standards might have changed in latter census year.

1.3 80-20 Male

"/disk/bulkw/sun.lee/matches/matches8020-1880notin1920.csv" stores files that are in 1880 but not in 1920. Most of them contain records born in asian countries and with small size except one file, containing white

male records fro other US possession. This file could be matchedif we have more detailed information about the birth place. "/disk/bulkw/sun.lee/matches/matches8020-1920notin1880.csv" stores files in 1920 but not in 1880. This file is interesting because it reflects the immigration trend into US. Most of the files in 1920 but not in 1880 are foreign born, containing 746,931 records in total, which indicates that between 1880 and 1920, people were immigrating into US from all over the world.

Similarly, for common race-BPL categories not matched in 80-20, the main reason is that the size of the race-BPL category in at least one of the years is small.

2. Matched Rates

Total male records is 53,886,871 in 1920, 62,104,510 in 1930 and 66,198,373 in 1940. For matched records after pruning, there are 18,508,302 in 20-30, 13,750,702 in 30-40 and 4,696,175 in 20-30-40. A rough estimate of 20-30 matched rate (after pruning) is 34.34% (18,508,302/53,886,871) and 20.77% (13,750,702/66,198,373) in 30-40. Moreover, about 25.37% (4696175/18508302) of 20-30 matched records are kept in 30-40 matched databases, which accounts for 34.15% (4696175/13750702) of the 30-40 male matched database. This in some extent reveals the dynamic change of the population. More interesting analysis involves analyzing the demographic and social economic characteristics of the 2-period snapshot matched databases and the 3-period snapshot matched databases.

2.1 20-30

"/disk/bulkw/sun.lee/matches/matches2030-matched_rate.csv" stores the matched rates of each race-BPL category in an ascending order. It is not hard to observe that almost all files with a matched rate higher than 35% are white male. And the matched rates are also relatively high using number of records in 1930 as denominator. This implies that the composition of the population of these race-BPL categories are relatively stable. In other words, mortality rates, names changed rates, transcription error rates and race changed rates of these race-BPL categories are relatively low compared to other files. Refering to the mortality records would provide us extra evidence for this conjecture. Visualizing the matched rates on map could also be an interesting exercise.

2.2 30-40

Similarly, "/disk/bulkw/sun.lee/matches/matches3040-matched_rate.csv" stores the matched rates of each race-BPL category in an ascending order. There are 34 files with matched rate higher than 35% in 20-30 combination. However, only 14 files have matched rates higher than 35% in 30-40 combination. The lower matched rate is also reflected in the overall matched rate. This is interesting given the comparability of 20-30 and 30-40 combinations, in the sense that both cases attempt to match records in adjacent census waves. My conjecture is that there may be more people immigrating in and out between 1930 and 1940 compared to 1920 and 1930. Also, mortality rate could be different due to wars and natural disasters. Further investigation could focus on the shift of population composition between adjacent census waves.

2.3 20-30-40

The above comparison on common race-BPL categories between adjacent census waves provides a crude indicator on how the population composition shifts across years, while linking 20-30 and 30-40 matched records provide us more detailed information on how the population changes across years. "/disk/bulkw/sun.lee/matches/matches203040-matched_rate.csv" keeps all race-BPL categories that are matched in 20-30 or 30-40.

2.4 80-20

Total records matched is 4,301,600. The number of the matched records is comparable to that of 20-30-40. As can be imagined, the matched rate would not be as high as those of latter years combinations. The highest matched rate is 32%. Similarly, those with high matched rates are usually white male.