Racial/ethnic self-identification: issues for demographic forecasting

Ethan Sharygin
Demographic Research Unit, California Department of Finance
ethan.sharygin@dof.ca.gov

SCAG 27th Ann. Demographic Workshop: June 13, 2016

Introduction

California is a racially and ethnically diverse state that demands a sophistication in handling issues related to racial/ethnic self-identification and demographic behavior. This brief introduces some of the innovations being made by the Demographic Research Unit of the California Department of Finance (DOF) in order to improve the ability of applied demographers to accurately model and forecast demographic outcomes in a multiracial, multiethnic population.

A central concern in running a population projection is accurately modeling and applying demographic rates. Gender, age, and race or ethnicity are the most commonly available traits for births and deaths for most of the history of vital statistical system in the US. While some behaviors may correlate better by education or nativity, it is much easier to attach a schedule of mortality and fertility rates to a population structured by age, gender, and race/ethnicity – characteristics that are widely reported and which either do not change or which change very predictably. Population forecasts are also expected to include an accurate description of the future race and ethnicity of the population, which can be affected by the way in which people already alive might alter their identities as well as the way that children not yet born will eventually self-identify when they take the Census.

Methods of classifying race/ethnicity

The categories used on the 2000 and 2010 Censuses allow up to 82 choices for racial/ethnic self identification– compare that to just 3 categories in the first Census. People have demanded the ability to identify themselves in the Census with more nuance than ever. Yet there are big unknown questions about interpreting the data. Crucially from the perspective of a demographer, are the categories predictive of actual differences in demographic behaviors or outcomes?

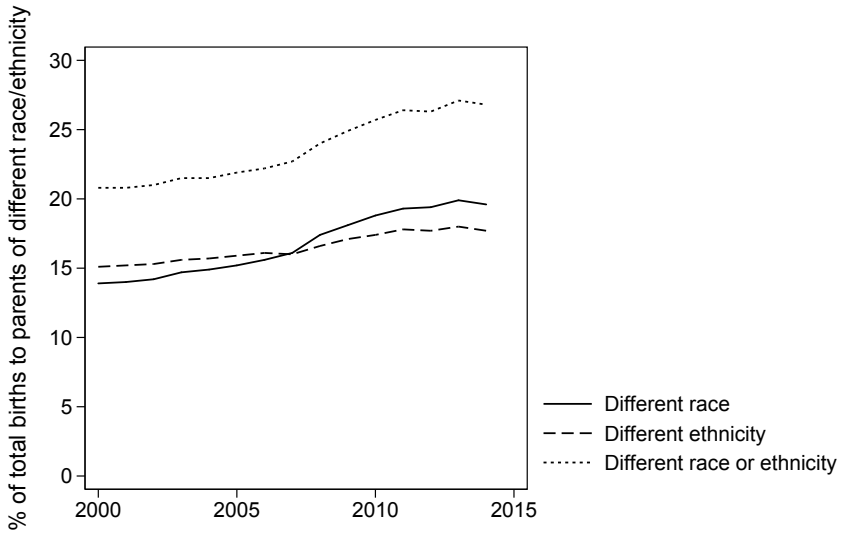
We embrace the change that allows individuals to report whichever race/ethnic categories they prefer. The approach we take at DOF is to probabilistically bridge people between the full 82 categories of identification to the narrower 5-category range used by the National Center for Health Statistics (NCHS), which we believe still captures essential differences in fertility, mortality, and migration propensities.

Race bridging algorithm: why and how to simplify analysis by race

Approximately 1 in 20 (4.6%) Californians are multiple-race, compared to 1 in 30 (3%) for the US as a whole. And this may continue to grow: 13% of children in California were born to couples where the mother and father identified differently by race, and fully 25% of births had parents either of different races or one Hispanic and one non-Hispanic parent.

The bridged race approach is based on the theory that demographic rates are more likely to be correlated with an individual's bridged race; in other words, that multiracial people have more in common with those of a particular other single race than to all other multiracial people.

Figure 1: Percent of births to parents of different race/ethnicity; California, 2000-2014



Source: California Department of Public Health, various years; author's calculations.

For example, how would one assess the likelihood of a 33 year old woman in LA county giving birth? We could look only at her location and age, or we could investigate other traits available in both the Census and the vital records, such as race/ethnicity. We find in the Census that she identified as non Hispanic and "American Indian (AI), Black, and Native Hawaiian or other Pacific Islander (NHPI)". We then check a third data source to see how often people of this mixed race background who are similar on the basis of age, sex, and geography reported using the 5-category race range.[1] Perhaps such a woman would have identified only as Black if given the old style Census form, and they would have been assigned the age specific fertility rates modeled by the previous outcomes for Black women in the forecast.

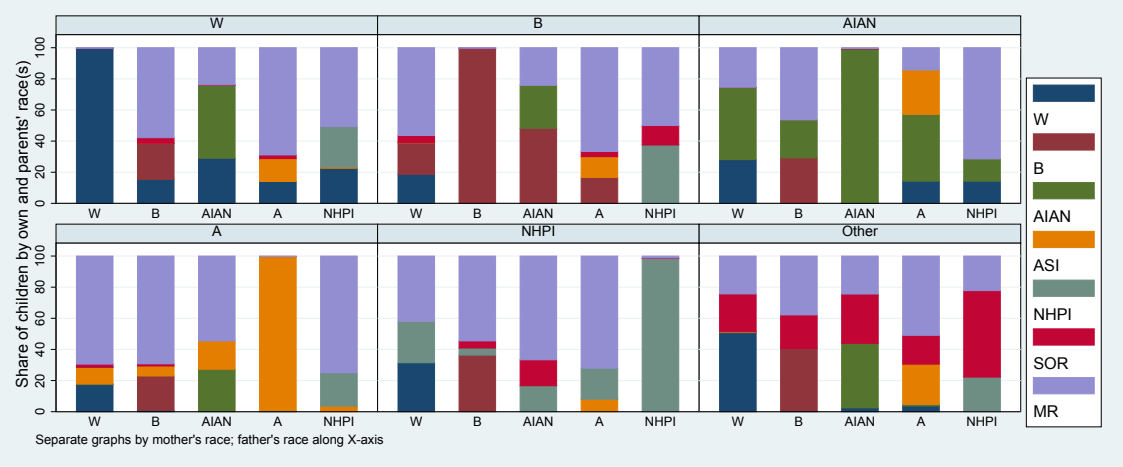
There are two drawbacks to the NCHS approach. Unfortunately, the model that NCHS uses is calibrated to data from 2001, and has not been updated since that time. Also, the NCHS approach identifies only the most likely single race that multi race individuals would choose and reassigns them to that single race when calculating mortality and fertility rates. To address some of the issues, the DOF forecast model considers not only the most likely category but the actual probability of reporting each category – which could hypothetically be something like 75% Black, 15% AI, 10% NHPI. We would then randomly draw a number between 1 and 100, and use that number to choose one of those categories. Rather than replacing a person's race in our projections dataset, we add the selected bridged race as an additional trait that is only used for analyzing mortality, fertility, and migration.

Determining child race from mother's traits

Since 2000, California has followed NCHS practice of reporting births according to the race of the mother. A child is not explicitly assigned to a race until time of death. Before that, the race of the child was determined according to various normative coding schemes that combined information on the mother and father. However, in the last 5 years, the proportion of records missing fathers' traits has been growing, and father's race/ethnicity has been missing from over 10% of birth records in California as of 2012. While the combined mother and father approach remains tenable after father's data are imputed, there are other problems. For example, basing child race on both parents' traits has results in vast over-estimation of the size of the multi-racial population, since in reality many children whose parents are from different racial/ethnic backgrounds in fact do not self-identify as multi-racial.

[1] The Minnesota Population Center has taken the NCHS model and produced Stata computer code to perform the race bridging on microdata files with the requisite covariates.

Figure 2: Percent of births to parents of different race/ethnicity; California, 2000-2014



Source: American Community Survey, various years; author's calculations.

Mothers traits alone are highly predictive of the racial identification of children, including the number of multi-race children. There are limited data from which to simultaneously assess parents and child race. The American Community Survey (ACS) reaches approximately 2% of US households each year, and contains a report on the race and birthdate of everyone in the household, as well as their relationship to the head of household. These questions can be used to compare the percent of children born to a given mother who are identified by a given race.[2]

As with race bridging, the approach DOF is taking is probabilistic, in which we examine the share of children who report each one of a possible 15 multiple race categories, together with Hispanic or non-Hispanic ethnicity. The share is modeled based on mother's own age, race, ethnicity, and the place where the mother lives.

Conclusion

An approach is presented in this brief that preserves a great deal of detail regarding racial self-identification, paving the way for future research on race and demographic behavior to be incorporated into the DOF forecasts. In most population projections, multi-race detail is discarded, and multi race individuals are either bridged to a single simplified race group or collapsed into a single multi-race group. This was sensible when mortality and fertility rates are based on state or county averages. In an effort to enhance the value of its projections, DOF has undertaken to estimate new mortality, migration, and fertility rates by county and race, rendering unsatisfactory the status quo treatment of the multi-race population. Instead, our projections now preserve detailed responses to the current form of the Census race/ethnicity question, while making an informed guess about what set of age-specific demographic rates should apply to each person in the population.

[2] A drawback of the ACS for this purpose is that one respondent provides the races and ethnicities of other household members. It would be preferable to examine datasets where parents and children were asked independently. Work is underway at DOF to analyze one such survey – the National Longitudinal Surveys (NLS).