PAA 2013 POSTER ABSTRACT

Title: Evaluation of the U.S. Census Bureau's County-Level Postcensal Population Estimates by Demographic Characteristics

Authors: Ben Bolender, Tiffany Yowell, Irene Dokko

BACKGROUND

The purpose of this project is to compare the results of the U.S. Census Bureau's official postcensal county population estimates series with 2010 Census counts by age, sex, race, and Hispanic origin. This is one aspect of a much larger project that has evaluated estimates of the population across national, state, county, and subcounty geographic levels. The primary focus here is on the distribution of these population estimates across age, sex, race, and Hispanic origin categories. We examine both simple differences between the population estimates and the Census counts and more complex "measures of accuracy" over a cross-classification of demographic characteristics and population size groupings.

The postcensal population estimates program calculates population figures for the nation, states, and counties by demographic characteristics on an annual basis. The basic principle is that we start with the previous decennial census counts (with modified race) and estimate forward primarily using administrative record inputs roughly based on the demographic balancing equation. In its simplest form, this means that population at a given time equals the population from the base (the previous census) plus births, minus deaths, plus net migration. National population estimates are produced using birth and death records, information on the military population, estimates of international migration, and group quarters information. State and county population estimates account for all of these components of change (both individually and through higher-level geographic controls) and additionally include the effects of domestic migration.

Our analysis is primarily based on the Vintage 2010 county characteristics data file (summed to higher levels of geography as necessary). This file contains April 1, 2010 population estimates for 3,143 counties for 86 categories of age (single years of age 0-84, 85+), 31 categories of race (White, Black, American Indian/Alaska Native, Asian, Native Hawaiian/Pacific Islander, and all combinations of these races), male and female, and two categories for Hispanic origin (Hispanic/non-Hispanic). This creates a dataset with roughly 33.5 million records, a majority of which have values of zero (since there may not be, for instance, any 83-year-old, Hispanic, White-Black-Asian people in a given county). This analysis will briefly present a summary of the results of the estimates evaluation for the population totals to provide reference for the discussion of results by characteristic.

For the purpose of this analysis, characteristic detail is collapsed into broad categories by separate characteristic detail (age groups, six major races, etc.). This eases a methodological problem related to cells containing zeros. Many of the measures of accuracy being analyzed include calculations with census or estimates populations as the denominator. This is an issue because division by zero is mathematically undefined, and therefore, impossible. Collapsing the categories dramatically reduces the

number of cells with zero values, thus making calculations more reasonable without having to drop larger numbers of geographic units or characteristic groupings from the analysis. Further, it allows us to see that apparent "inaccuracies" in the population estimates may be related to known difficulties in estimating small population groups.

METHODS

The population estimates are compared with 2010 Census values using a variety of "measures of accuracy." In this portion of the project, we focus on the following three meaures.

Mean Absolute Percent Error (MAPE)=((∑ ((|Estimate – Census|)/ Census))/N)*100

The goal of the mean absolute percent error (MAPE) is to provide a relative measure of error. It ranges from zero to positive infinity and represents the average error across cases, regardless of sign. Also, this is one of the most commonly used techniques available, and is therefore readily understood and easily accessible.

Mean Algebraic Percent Error (MALPE)=((∑((Estimate – Census)/Census))/N)*100

Similar to the MAPE, the mean algebraic percent error (MALPE) looks at the relative size of differences across counties. However, its purpose is to identify systematic bias and provide an alternative for a relative measure of error. Its main value is that it preserves the sign of the error, allowing us to assess whether the population estimates were generally higher or lower than the census count.

Dissimilarity Index=(∑|(% in County by Group in Census 2010 - % in County by Group in Estimates)|)*0.5

While the MAPE and MALPE are specific to the group being measured, the dissimilarity index allows us to evaluate an entire distribution in a single measure. The dissimilarity index ranges from 0 to 100 and can be interpreted as the percent of people who would need to change categories in order for the distribution to match between two datasets. Whereas we use the MAPE and MALPE as summary measures of counties at the national level, the dissimilarity index can be computed for each county individually, allowing us to see how our accuracy varies over space.

We used some of these measures in our previous evaluation work, and they continue to be relevant when looking at counties by characteristic detail. Using similar measures allows us to compare findings between the characteristic groups and the total. Further, the measures are not geographically specific. Some methods to evaluate accuracy break down when used in situations where the geographies are not complete. Each of these four measures can be considered separate of the geographic distribution of the units themselves. Regardless of how many counties contain zero values for particular groups (e.g., Native Hawaiian and Other Pacific Islander), the four measures are still statistically relevant.

RESULTS

We start with results of the estimates evaluation for population totals. Overall, the population estimates were very close to 2010 Census counts of the total population. At the national level, the population estimates only differed from the Census by -0.1 percent in the published population estimates or about -0.3 percent for population estimates created purely from administrative records (our "Pure ADREC" series, not taking into account the effects of challenges and special censuses). For counties, the mean absolute difference was only 3.1 percent for the published population estimates or about 2.9 for the "Pure ADREC." This provides a general backdrop for the more focused discussion on the distribution of measures by characteristic group.

Next, we show measures across counties at the national level for collapsed categories of demographic detail. The first part of this section covers summary measures across individual characteristic groups (e.g., age or Hispanic origin by themselves). The second part presents similar measures with these same broad groupings cross tabulated (e.g., age group by Hispanic origin). This analysis also uses group size thresholds to examine the impact on the measures of using very small groups.

For example, the MAPE for people ages 18 to 24 can be as high as 12.5 percent if we look at all counties. This is not surprising as this age category is extremely hard to estimate due to the transitory nature of people at that point in life (college, military, and movement away from childhood homes). Limiting the analysis to areas of 500 or more people in that group drops the MAPE to about 9.3, and looking at counties with 1,000 or more people lowers it further to about 8.2 percent. This size relationship becomes more apparent when looking at race and Hispanic origin. When looking at estimates of the population of Hispanics, all counties show a MAPE of about 34.2 while estimates in counties with a group population of 1,000 or more are only different from 2010 Census counts by about 10 percent.

This analysis demonstrates that the Census Bureau's Population Estimates program is very accurate when it comes to the total population. However, there are larger differences when looking at our population estimates by demographic characteristics. On some measures of accuracy, the population estimates may appear quite different from Census counts. However, a more detailed analysis shows that the main difference lies in the estimates of small populations, not differential measurement of various characteristic groups.

The poster itself presents the analysis of selected measures of accuracy for collapsed age, sex, race, and Hispanic origin categories. We divide the population into six age categories (less than 10, 10-17, 18-24, 25-64, 65-84, and 85+) and eight race/Hispanic origin categories (non-Hispanic White alone, non-Hispanic Black alone, non-Hispanic American Indian or Alaska Native alone, non-Hispanic Asian alone, non-Hispanic Native Hawaiian or Other Pacific Islander alone, non-Hispanic Two or More races, Hispanic Two or More Races, and Hispanics of One Race). We both examine accuracy of the population estimates for separate groups and as a whole through the distribution of characteristics. Further, we point out some interesting case studies for discussion and future research.