Compounding the Error? Evaluation of the U.S. Census Bureau's County Population Estimates

Michael E. Cline
Michael.cline@osbm.nc.gov

Extended Abstract

Population estimates play an important role in public policy. Population estimates are used to track trends after a decennial census and serve as a basis for population projections or other related forecasts needed for transportation, school, and other infrastructure improvements. On an annual basis, approximately $590 billion or $1,838 per person in federal funds are distributed to states and local communities for various programs based indirectly on the decennial census - primarily through formulas which use post-census population estimates directly or indirectly through population based surveys weighted to the post census population estimates (Reamer 2017). States may also use Census Bureau population estimates for their own formula funding programs. For example, in North Carolina, approximately $1.5 billion in state funds are distributed on an annual basis using population estimates prepared by the North Carolina State Demographer which partially rely on the Census Bureau county population estimates. The extensive literature on the evaluation of population estimates have helped improve population estimation methods and have consistently shown that, in general, estimation errors are inversely related to the population size of the areas being estimated or the rate of population change of the area estimated (Hoque 2010; Smith and Cody 2004; Swanson and Tayman 2012; Swanson, Tayman, and Barr 2000). In other words, the larger the population of the area estimated, the closer the estimate will be to the actual population and areas with little change will be easier to estimate than those with rapid growth or decline. Most evaluations use standard methods to evaluate the accuracy and bias of the method or methods used to estimate the population based upon standard classifications of county (or other geographic area) population size or degree of population change. However, these evaluations tell us little about any other biases that may result in areas being consistently over- or under-estimated during the post census period. Building on the work of Judson, Popoff, and Batutis (2002), this paper evaluates the Census Bureau's 2000 and 2010 population estimates to understand how selected characteristics of counties may lead to the over- or under-estimation of county population. An understanding of these biases (if any) are important for improving the methods used for population estimates.

Data

In this paper, two sets of population estimates are evaluated: those produced to estimate the population for 2010 and those produced to estimate the population for 2000. For the evaluation of the 2010 population estimates, the non-adjusted population estimates will be used. These estimates were prepared using the administrative records method in place in the post 2000 period and did not include any adjustments made to county population estimates based upon special censuses or local challenges. With minor adjustments in data sources and methods these population estimates use the same general methods as those used in the post-2010 census.

- U.S. Census Bureau, Alternative Vintage 2010 Population Estimates and 2010 Census Counts for the United States, States, Counties, Puerto Rico, and Subcounty Areas

The estimates prepared for the evaluation after the 2000 Census are used to evaluate the population estimates for that year.

- U.S. Census Bureau, 1990 Based July 1, 2000 Estimates Used for Estimates Evaluation

In addition to the evaluation datasets, this research will use county based characteristic data from the following datasets:

- 2010 U.S. Decennial Census, Summary File 1
- 2000 U.S. Decennial Census, Summary File 1 and Summary File 3
- 2008-2012 American Community Survey
- USDA ERS County Typology Codes

Research Methods

The first part of this research will involve an evaluation of the population estimates using standard measures to include Mean Algebraic Percent Error (MALPE), Mean Absolute Percent Error (MAPE), Percent Positive Errors (%POS), and percent absolute errors less than 5% and above 10%. Counties will be classified by size and rate of population change as well as on other characteristics such as USDA's persistent poverty counties or recreation counties. Then, using the Algebraic Percent Error (ALPE) as the dependent variable, a range of county characteristics will be evaluated to understand the level of bias (if any) for each of the characteristics relative to the population estimates.

The dependent variable is the Algebraic Percent Error (ALPE), which is:

ALPE = 100 * (POPEST – COUNT)/COUNT where,

POPEST = 2010 (or 2000) date adjusted population estimate, and

COUNT = 2010 (or 2000) Decennial Census Count.

A positive value means that the population estimate was larger than the actual census count (the population was over estimated), while a negative value means that the population estimate was lower than the actual census count (the population was under estimated).

Control variables include population size in 2010 (or 2000) and percent population change from 2000 to 2010 (or 1990 to 2000). The independent variables are described in Table 1. The expected direction of relationship to the dependent variable is also shown for each characteristic. A positive bias means that the particular characteristic is expected to lead to an over-estimation of the population while a negative bias means that the characteristic will result to an under-estimation of the population. These variables were chosen based upon the author's experience in estimating population as well by the work of Judson, Popoff, and Batutis (2002) that hypothesized biases due to limitations in administrative records data.

Table 1. Variables Used in Analysis

Variable Name	Definition	Expected Bias
Dependent		
ALPE(x)	Algebraic Percent Error for given year x	NA
Controls		
POP(x)	Decennial census count for given year x	
POPSIZE(x)	Population size for given year x	
Independent		
NATINC	Natural increase / census population	+
RURAL	% population living in rural areas	-
NOHOSP	No hospital in county = 1	-
RURLHO	Interaction RURAL*NOHOSP	-
RECR	USDA ERS Recreation County = 1	-
PCTLOED	% population 25-64 without high school diploma/GED	-
PCTINST	% population living in institutional group quarters facilities	-
PCTDORM	% population living in college dorm	-
PCTMIL	% population in military	-
PCTPRIS	% population living in a prison	+
PCTFB	% population foreign born	-
BLACK	% population black	-
HISP	% population Hispanic	-
IND	% population American Indian/Alaska Native	-
PCTPOV	% population in poverty	+
MEDHHINC	Median household income	+
PCTMED	% of population 65+ with Medicare coverage	+
MEDAGE	Median Age	+

References

Hoque, Nazrul. 2010. "An Evaluation of Small Area Population Estimated Produced by Component Method II. Ratio-Correlation and Housing Unit Methods for 1990." *The Open Demography Journal* 3:18–30.

Judson, D. H., Carole E. Popoff, and Michael J. Batutis. 2002. "An Evaluation of the Accuracy of US Census Bureau County Population Estimates." *Statistics in Transition* 5(2):205–35.

Reamer, Andrew. 2017. *Counting for Dollars 2020: The Role of the Decennial Census in the Geographic Distribution of Federal Funds*. Washington, D.C.: George Washington University.

Smith, Stanley K. and Scott Cody. 2004. “An Evaluation of Population Estimates in Florida: April 1, 2000.” *Population Research and Policy Review* 23(1):1–24.

Swanson, David A. and Jeff Tayman. 2012. “Evaluating Estimates.” Pp. 267–302 in *Subnational Population Estimates*. Vol. 31, *The Springer Series on Demographic Methods and Population Anlaysis*. Springer Science+Business Media.

Swanson, David A., Jeff Tayman, and Charles F. Barr. 2000. “A Note on the Measurement of Accuracy for Subnational Demographic Estimates.” *Demography* 37(2):193–201.