Subnational Estimates of Net Coverage Error for the Population Aged 0 to 4 in the 2010 Census

Heather King
David Ihrke
Eric Jensen

Population Division
U.S. Census Bureau

Presented at the Annual Meeting of the Population Association of America,

Denver, CO, April 26-28, 2018.

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

ABSTRACT

Young children aged 0 to 4 had an estimated net undercount of -4.6 percent in the 2010 Census compared to a 0.1 percent overcount for the total population. Net coverage error for these cohorts is estimated using Demographic Analysis (DA). DA uses historical vital records and data on international migration to produce estimates of the population. The 2010 DA estimates were produced at the national level; therefore, the data cannot be used to estimate net coverage error for states or counties. In this paper, we produce subnational DA estimates of the population age 0-4 using vital records, international migration data, and domestic migration rates at the state and county levels. The results will show the geographic areas where young children had the highest estimated net undercount in the 2010 Census.

Introduction

Young children aged 0 to 4 had an estimated net undercount of -4.6 percent in the 2010 Census compared to a 0.1 percent overcount for the total population. Net coverage error for these cohorts is estimated using Demographic Analysis (DA). DA uses historical vital records and data on international migration to produce estimates of the population. The 2010 DA estimates are essential to research on the undercount of young children, but these data are limited. The DA estimates were only produced at the national level and therefore subnational estimates of net coverage error for young children are not available.

Understanding patterns of geographic areas with the largest undercounts for young children would enable the Census Bureau to design strategies and operations to improve the count for this population in the 2020 Census. While several studies have used Vintage 2010 Population Estimates to measure the undercount of young children at the state and county levels, these data are not appropriate for measuring coverage. It is clear that subnational DA estimates are needed to evaluate the coverage of young children at the state and county levels.

In this paper, we leverage the strength of several administrative data sources, in conjunction with survey data, to develop estimates measuring net coverage error in the 2010 Census of young children aged 0 to 4 at the state and county levels. Using vital records on birth and death data, domestic migration rates, and data on international migration, we produce state and county DA estimates of the population aged 0 to 4 as of Census Day (April 1, 2010). Next, we compare these estimates to counts from the 2010 Census to calculate net coverage errors. We also compare our subnational coverage error estimates to those identified using the Vintage 2010 Population Estimates[1]. We then use spatial and cluster analysis to highlight patterns in the geographic distribution of coverage errors for young children by demographic and housing characteristics.

Background

The 2010 Census had an estimated net undercount for young children aged 0 to 4 of -4.6 percent, which was higher than for any other age group. Research on the undercount of young children has found strong relationships between race, ethnicity, and household structure and the coverage for this population (U.S. Census Bureau 2017a, 2017b, 2017c). Research has also found that coverage of young children in the 2010 Census may vary by state and county. For example, considerable differences were identified between the Vintage 2010 Population Estimates (V2010) and 2010 Census counts for young children across states and counties (O'Hare 2015, U.S. Census Bureau 2017).

In the 2010 Census, O'Hare found that 9 out of the 10 most populous counties showed an estimated net undercount exceeding -10 percent, more than twice the national number of -4.6 percent. The O'Hare analysis also showed that about 77 percent of the estimated net undercount occurred in the 128 largest counties (O'Hare 2015). Other research has shown large differences in the estimated net undercount in New York and Illinois between New York City and the rest of the state and Cook County (Chicago) and the rest of the state (U.S. Census Bureau 2014).

[1] Here we use a series of research estimates released in March 2012 that were intended to evaluate the accuracy of the Census Bureau's annual population estimates. These estimates do not incorporate special census or challenge results. For more details, see https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates.html

Research showing state and county estimates of the undercount of young children in the 2010 Census use the V2010 estimates because the 2010 DA estimates were only produced at the national level. In addition, the V2010 estimates for young children are similar to DA because the estimates for those cohorts are not based on the prior census, but are developed primarily using birth records. However, domestic migration estimates in V2010 were developed using some data from Census 2000. DA is an official method used by the U.S. Census Bureau to measure coverage in the census. Comparisons between the Population Estimates and the decennial census, referred to as the "error of closure," are used to evaluate the quality of the estimates and not the census counts.

In this paper, we produce subnational DA estimates of the population age 0 to 4 and use them to estimate net coverage error for young children at the state and county levels. We expect to provide a clearer picture of the subnational distribution of the undercount of young children beyond what has been done using the V2010 estimates for four main reasons. First, the V2010 estimates used projected birth data to develop estimates of 0 and 1 year olds, but we now have full birth data for all ages. Next, we are making improvements to how we process the vital records. Furthermore, we have developed domestic migration assumptions tailored to ages 0-4. Finally, we are incorporating data from Mexico on young children born in the United States, but living in Mexico at the time of the 2010 Census (Jensen, Benetsky, and Knapp 2018).

Data and Methods

Subnational DA Population

To examine the net coverage error of young children in the 2010 Census, we produce a county level Demographic Analysis (DA) series as of Census Day (April 1, 2010). We use a cohort component method for births, deaths, and domestic migration and a residual stock method for estimates of net international migration (NIM).

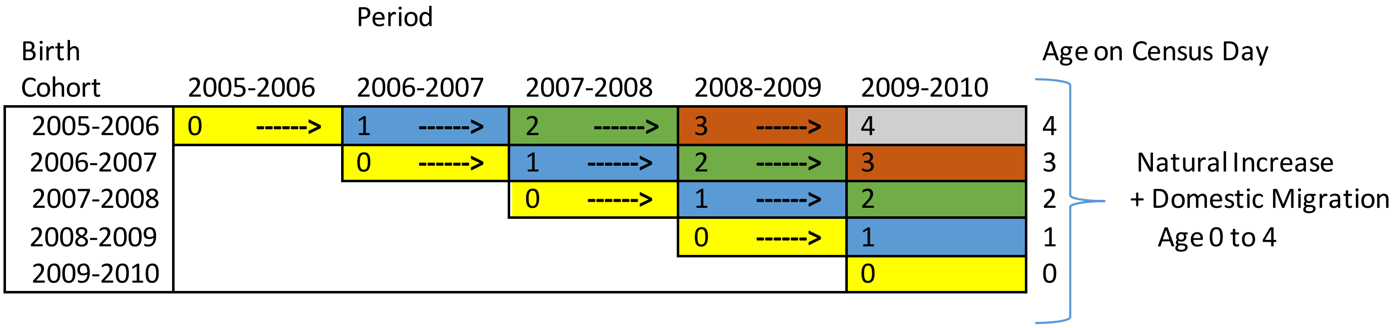
The cohort component portion of the estimate spans 5 annual periods that cover April 1, 2005 to March 31, 2010. Figure A on the following page demonstrates the cohort component portion of our Subnational DA process and shows how we build the population age 0 to 4 on Census

Day. Each row represents a county birth cohort, each column is a period, and the color codes refer to age during that period. The starting population is the cohort of births over the first period (2005-2006) and is shown in the yellow cell in the first row in Figure A. The starting population comprises children who are age 4 on Census Day. We subtract deaths that occurred over the first period to the starting population and account for domestic migration between counties to obtain the population age 1 at the start of the second period.

In the second period (2006-2007) we add new births, subtract deaths to the 0 and 1 year olds, and account for domestic migration. In other words, we account for natural change over the period and distribute domestic migrants by age and county. We continue this process until April 1, 2010.

On Census Day, we add the NIM portion of the estimate by county and age to our estimates of cumulative natural change and domestic migration. We then sum over age to obtain county total estimates for young children, and then sum counties to get estimates for states.

Figure A. Subnational DA Estimates Cohort Component Method: Births, Deaths, and Domestic Migration



		Period				
Birth Cohort	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010	Age on Census Day
2005-2006	0 ------>	1 ------>	2 ------>	3 ------>	4	4
2006-2007		0 ------>	1 ------>	2 ------>	3	3
2007-2008			0 ------>	1 ------>	2	2
2008-2009				0 ------>	1	1
2009-2010					0	0

The following sections focus on the input data and methodology used to develop the components of our Subnational DA series: births and deaths, domestic migration, and international migration. See Tables 1a and 1b at the end of the paper for the full inventory of input data used to produce the Subnational DA estimates and the following analysis of net coverage error. After a description of the input data, we discuss measures of net coverage error, the methods used to analyze net coverage error, results and major findings, and then finish with a note on limitations and next steps.

Births and Deaths

The 2010 national DA estimates relied primarily on individual administrative records on births and deaths from the National Center for Health Statistics (NCHS). The subnational DA estimates presented in this paper are based on these same data on vital events, with some processing differences and additional data sources.

In the subnational DA series, we supplement the NCHS birth series with the following additional sources from state governments:

1. Publicly available county births by year from state vital statistics/public health offices (2005-2007)
2. Federal-State Cooperative for Population Estimates (FSCPE) county births by year for years 2008-2010

We use county data from the above state sources to improve the county distribution of births in the NCHS vital statistics data. Theoretically, these state data and the NCHS data on births should be identical, since state vital statistics offices submit these same data on all births to their residents and those that occur within their state to NCHS. In practice, the timing of reporting and varied levels of cooperation between the states in information sharing sometimes gives different results. NCHS reconciles all births by mother's county of residence and county of occurrence of the birth. This is not always straightforward with states that have complex geographical boundaries. Virginia is an example, a state with 38 county equivalent independent cities.

The Federal-State Cooperative for Population Estimates (FSCPE) is a cooperative between the Census Bureau and state governments formed to improve the quality of the Census Bureau's annual population estimates and to facilitate data sharing and methods. The group is comprised of Census Bureau staff and state demographic experts on behalf of their respective governor's offices[2]. Participating state FSCPE members submit annual data on county vital events to supplement the Census Bureau's Population Estimates Program (PEP) state and county estimates. These data improve the county distribution of vital events provided by the National Center of Health Statistics, because state FSCPE members provide valuable demographic review, feedback, and local expertise.

In the Census Bureau's annual population estimates, FSCPE data are used to estimate the county distribution of vital events by state, since local demographers tend to have better knowledge/data at this level compared to NCHS. However, we preserve NCHS state totals in the final estimate since NCHS compiles all final births and can reconcile births that occur to residents out of state. In the subnational DA estimates presented here, we follow this procedure. Thus, the final birth estimates for counties over each period follow the county distribution supplied by state data but sum to the state totals from NCHS.

Data on annual county births from FSCPE members span from 2008 to as recent as 2017. To fill in the county distribution from state data sources for the years 2005-2007, we use publicly available data from state vital statistics/public health office websites, when available.[3] The county birth data from state websites show high agreement with the data submitted by FSCPE members in the years following 2007. We found that around 90 percent of county totals from state websites were equal to the number of births supplied by FSCPE members. Another 7 percent were within +/- 10 births.

[2] See the FSCPE website for more details about its purpose and function in the Population Estimates Program at https://www.census.gov/programs-surveys/popest/about/fscpe.html.

[3] We use NCHS data for Nevada, Louisiana, Montana, Minnesota, Maine, and Vermont for some years.

In our Subnational DA birth series by county, we append the FSCPE series (2008-2010) to the state public series (2005-2007) to obtain a county distribution of births from state data sources. We then reconcile this series with NCHS births–one that preserves the county distribution of births within each state and keeps the state total tabulated by NCHS. See Appendix A for more details on how this is accomplished.

Annual county deaths are estimated solely from NCHS data, because the years of data and detail available on age of decedent at the county level vary widely on state public websites. An analysis of infant deaths in states with the highest discrepancies between NCHS county births and state births showed minor differences between deaths from the two sources. At any rate, the death component is overwhelmed by the large number of births for these ages, so small differences between NCHS and state death data are acceptable.

After reconciling NCHS and state data on county births, we produce county birth estimates by period and death counts by county and age of decedent for each period to build the population from April 1, 2006 to March 31, 2010.

Domestic Migration

After accounting for births and deaths to each cohort for every period, we distribute domestic migrants to counties using a combination of out rates and in proportions of internal migration based on administrative data from the Internal Revenue Service (IRS) and Social Security Administration (SSA). We use person level data from both sources to produce rates, where tax return data provide residence information and the SSA data supply date of birth. We combine these data to obtain a series of out migration rates and in migration proportions by single year age for each county and period.

To derive rates, we match person level IRS tax return data over two tax years by an identifier called the Protected Identification Key (PIK)[4]. We use the address information provided on the IRS tax return to determine county migration status. A tax return includes the filer's address information, so we use the ZIP code to ascertain county of residence. All individuals that appear on the same tax return share the same address. Thus dependent children, our population of interest, are assigned the address of the filer who claimed them. A record in the tax data that changes county from one tax year to the next is tallied as a county migrant. When the county stays the same across two years, the record is tallied as a non-migrant.

Next we match the IRS migration universe by PIK to the SSA NUMIDENT file, a database of all Social Security Numbers ever assigned, to append age to each record. With age appended, we can develop estimates of migration by single year of age for each migration period, where age is calculated as of the beginning of the period.

To calculate migration rates, we tally the number of county in migrants, out migrants, and non migrants for each county, age, and period determined by the tax data. We then develop a series of out rates, as follows, for every county *i*, at age *j*, over the period *k*:

$$out\ rate_{i,j,k} = \frac{out\ migrants_{i,j,k}}{non\ migrants_{i,j,k} + out\ migrants_{i,j,k}}$$

The out-rate at age *j* in county *i* is the proportion of individuals in the tax data *j* years old at the beginning of the period who moved out of the county.

To estimate in migration, we develop a proportion for each county, single year age, and period that allocates the national pool of migrants by age to a destination county, as follows:

[4] A PIK is an identifier on person records that protects confidentiality and enables matching across multiple administrative data sources. For more details, see https://census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-02.pdf

$$in\ proportion_{i,j,k} = \frac{in\ migrants_{i,j,k}}{\sum_{i=1}^{3,143} migrants_{i,j,k}}$$

For example, the in proportion for County A at age 1 in period 3 is the share of all 1-year-old migrants in period 3 who migrate to County A.

To create estimates of domestic migration in the Subnational DA series, we apply the out rates by single year of age to every county in each period, sum all county out-migrants to create a national pool of migrants, and then allocate the out-migrants to their destination counties by age based on the in-proportions derived from the IRS/SSA data.

International Migration

After accounting for natural change and domestic migration over all 5 periods between April 1, 2005 and March 31, 2010, we use data on international migration to complete the DA estimate of the population aged 0 to 4 for each county. The estimates of net international migration (NIM) for the 2010 DA national series were developed using data from the ACS and other sources. We used the totals, by age, from the 2010 DA estimates (Revised 2012 Series) for the foreign-born immigration, foreign-born emigration, net migration between the United States and Puerto Rico, and Born Abroad of U.S. Citizen Parents components.

In this study, the net native migration component was updated using data from Mexico on young children born in the United States but living in Mexico at the time of U.S. 2010 Census. Specifically, we used data from the National Survey of Occupation and Employment (ENOE), which is a monthly Labor Force Survey conducted in Mexico. The ENOE has a sample of approximately 40,000 households.

County level estimates of NIM were produced by distributing the national totals for each component to counties using the 2006-2010 5-year ACS file. The estimate of U.S.-born migration to Mexico was distributed to counties using information from the vital records on county of residence for births where the mother's place of birth was Mexico.

Net Coverage Error Analysis

Net Coverage Error

We use the Subnational DA series as a benchmark to study the net coverage error of young children in the 2010 Census. We define net coverage error as the percent difference of the Census 2010 population age 0 to 4 on Census Day from the population obtained from our Subnational DA estimates. We summarize the net coverage error of young children by state and county with the Mean Algebraic Percent Error (MALPE) and Mean Absolute Percent Error (MAPE). The MALPE used here is defined as the average net coverage error over analysis groups (which can be geographical units or demographic groups). The MALPE is calculated as follows:

$$MALPE = 100 * \frac{1}{n}\sum_{i=1}^{n}\frac{Census\ 2010 - Subnational\ DA}{Subnational\ DA}$$

where the Subnational DA estimates are the benchmark series over *n* units. The MAPE describes the average net coverage error in census counts from our Subnational DA estimates similarly, but we focus on the magnitude of difference instead of direction by taking the absolute value of the differences. We calculate the MAPE as follows:

$$MAPE = 100 * \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Census\ 2010 - Subnational\ DA}{Subnational\ DA}\right|$$

Net coverage errors that are negative are described as "undercount" in the census, while positive net coverage errors are described as "overcount."

We further analyze MAPEs and MALPEs by county population size, difference categories, and rankings.

In addition to numeric summaries, we also show a series of maps that give the geographic distribution of net coverage error of young children in Census 2010. Additionally, we compare the net coverage error patterns identified in this study to those identified by the Vintage 2010 Population Estimates.

Cluster Analysis

We end the analysis with an examination of net coverage errors by population and housing characteristics by using a cluster analysis to group together demographically similar counties into mutually exclusive groups. We then summarize the net coverage error by cluster to examine undercount by combined demographic and housing characteristics.

We compiled indicators on the demographic composition and housing characteristics (mostly from Census 2010) for each county to form a typology of demographically similar counties using a k-means cluster analysis. Grouping the counties in this way allows us to examine net coverage error by multiple, related demographic and housing characteristics. See Appendix B for details on the method used to obtain the clusters.

We identified 7 clusters:

1. **Majority Black and White**
2. **Majority American Indian/Alaska Native**
3. **High Hispanic proportion**
4. **Immigrant destinations**
5. **Prison and military counties**
6. **Average characteristics**
7. **Mostly White**

The defining characteristics of each cluster should be obvious from their labels, but we will expand on the description of each cluster later in the report when we discuss net coverage error.

Results

The 2010 Census enumerated 20,201,362 young children under 5, while the middle series of the revised 2010 DA national estimates showed approximately 21,171,000 (U.S. Census Bureau 2014). In this work, we increase the native born emigration component based on Mexican Census results in 2010, so we estimate 21,015,226 young children on Census Day. Consequently, we reduce the national undercount of children identified in the 2010 DA series downward by about 150,000 kids.

Our national number (Subnational DA summed over all counties) falls between the Census count and the revised DA official estimate. At the national level, we identified an undercount of 813,864 (-3.9 percent). While not as high as the undercount identified in the revised DA official estimates (-4.6 percent), the undercount of children seen here is substantial and higher than all other age groups (U.S. Census Bureau 2014). This update to the national net coverage error is seen in states and counties. Following is an analysis of state and county level patterns of net coverage error, using MALPEs, MAPEs, mapping, and cluster analysis and a comparison of net coverage error patterns using our Subnational DA series and the Vintage 2010 Population Estimates.

State Level Analysis

Using our Subnational DA series as the benchmark, the net coverage error of young children in the 2010 Census at the state level ranged from a high undercount of -6.2 percent seen in **Florida** to an overcount of 0.4 in **Idaho**. The Mean Algebraic Percent Error (MALPE), an average measure of net coverage error across all states, was -3.2 percent. The states with highest net coverage errors were generally in the **South**.

The 10 states with the highest net coverage error were **Florida** (-6.2), **Mississippi** (-6.1), **Delaware** (-5.7), **Georgia** (-5.6), **Virginia** (-5.4), **Alabama** (-5.4), **New Jersey** (-5.2), **Texas** (-5.0), **Maryland** (-5.0), and **West Virginia** (-4.7).

The state patterns of net coverage error were similar to those seen in the Vintage 2010 analysis, with two key differences:

1. net coverage errors were distributed more uniformly across the nation in the Subnational DA series compared to Vintage 2010, and
2. highest undercounts using the Subnational DA series were found in the **South** instead of the **West**.

The MALPE for states using the Vintage 2010 series was -3.2 percent, comparable to that seen in the Subnational DA series (rounded to -3.2 percent). Of the 50 states, both the Subnational DA and the Vintage 2010 series measured an undercount in 45. See Table 2 for the high level overview of the differences between Subnational DA and Vintage 2010 net coverage errors by state.

Both series measured net coverage error in nearly every state, but the Subnational DA series found slightly increased net coverage errors in 27 states and reduced the extreme levels seen in the Vintage 2010 series. See Figure 1 for the relationship between state level net coverage error using Subnational DA (vertical axis) and that using the Vintage 2010 series (horizontal axis).

In Figure 1, the state net coverage error identified by the Subnational DA series is plotted versus the same from the Vintage 2010 series. Points that fall on the Y=x black reference line indicate perfect agreement in net coverage error between the series. In general, there is high agreement as the points follow closely along the Y=x line. The red points below the reference line are the 27 states where both series measure a net undercount, but the Subnational DA series found a higher undercount. However, because the points are relatively “close” to the line, the increase in net undercount identified by the Subnational DA is slight.

On the other hand, the green points above the reference line in Figure 1 represent the states that had a higher undercount using the Vintage 2010 series. We see that these points show more dispersion and appear to be “further” from the reference line compared to the red points below the line. In other words, the Subnational DA series reduced the net undercount in the more extreme cases identified by Vintage 2010. For example, the Subnational DA series reduced the net coverage error identified in **Arizona** by the Vintage 2010 series from -10.0 to -4.2.

The Subnational DA found the highest undercounts in the **South**, while the Vintage 2010 analysis generally showed the highest state level net coverage errors in the **West** (**Arizona**, **California**, **Texas**, and **Nevada**) and in states with large Hispanic populations (**Florida**, **Georgia**, and **New York**).

The Vintage 2010 series found that 10 states exceeded -5 percent in net coverage error (compared to 9 using this same threshold for the Subnational DA series): **Arizona** (-10.0), **California** (-7.5), **Florida** (-7.5), **Texas** (-7.3), **Georgia** (-7.1), **Nevada (**-6.6), **Illinois** (-5.6), **New Mexico** (-5.4), **Delaware** (-5.4), and **New York** (-5.3). See Table 3 for the complete list of net coverage errors of states using both the Subnational DA series and Vintage 2010.

Compared to the Vintage 2010 series, the net coverage errors using the Subnational DA were similar but distributed more uniformly across the nation with fewer outliers in net coverage error (as shown in Figure 1) and identified the highest undercounts in the **South** instead of the **West**.

County Level Analysis

Table 4 summarizes the net undercount of young children in the 2010 Census at the county level using the Subnational DA and Vintage 2010 estimates. Overall, the Subnational DA estimates measure more undercount than the Vintage 2010 estimates, while also reducing the number of counties with extreme values. This pattern of difference mimics that seen at the state level.

The percentage of counties with 10 or more percent net undercount declined from 8.9 percent in the Vintage 2010 estimates to 6.5 percent in the Subnational DA. The largest difference between the two series was in the number of counties that showed a moderate (at least 1 but less than 10 percent) undercount for young children. In the Subnational DA estimates, 55.2 percent of counties showed a moderate undercount compared to 37.2 percent in the V2010 estimates. The percentage of counties with close to full coverage (less than +/-1 percent) was similar across the two series. The percentage of counties with either a moderate (at least 1 but less than 10 percent) or high (10 percent or more) overcount was higher in the V2010 estimates.

Figure 2 maps the spatial distribution of the net undercount of young children in the 2010 Census using the Subnational DA estimates in the categories reported in Table 4. The high net undercount counties are found in the Mississippi Delta, Appalachia, border counties in **Texas**, and counties with large American Indian/Alaska Native populations. The counties with a moderate undercount for young children are dispersed throughout the United States. Counties with full coverage, moderate, and high overcounts were concentrated in New England, the Great Plains, and rural counties in the **Wes**t.

Table 5 shows the mean absolute percent error (MAPE) and mean algebraic percent error (MALPE) for the percent net undercount and error of closure[5] for young children in the 2010 Census. For all counties, the MAPE between the Subnational DA estimate and the Census 2010 count for young children was 4.9 compared to 7.4 for Vintage 2010. This indicates that overall the Subnational DA estimates are closer to the census counts than Vintage 2010. However, the MALPE for all counties for the Subnational DA was -1.7 compared to 1.2 for Vintage 2010, which means that the Subnational DA series identified a higher undercount overall at the county level.

[5] Comparisons between the Population Estimates and the decennial census, referred to as the "error of closure," are used to evaluate the quality of the estimates and not census counts.

We also calculate MAPEs and MALPEs by county size (Table 5). In both the Subnational DA and Vintage 2010 estimates, there was a curvilinear relationship between MAPE and county size, with the largest MAPE values in the smallest and largest counties. The MAPE for the Subnational DA estimates was either lower or equal to the MAPE for the V2010 estimates for all size categories. For the smallest counties (less than 5,000), the MALPE in the Subnational DA and Vintage 2010 estimates was positive, indicating an overcount in those counties. For all other size categories, the MALPE for the Subnational DA was negative. The largest counties (500,000 or more) had the largest negative MALPE in both the Subnational DA and V2010 estimates.

Table 6 reports the similarities and differences between net coverage error estimated using the Subnational DA and the V2010 estimates. Nearly half of all counties (45.2 percent) showed an undercount in the census for young children using both sets of estimates. Conversely, 23.9 percent of counties showed a net overcount. Some counties flipped from a net undercount to a net overcount and vice versa when we compare the two sets of estimates. In 24.6 percent of counties, the Subnational DA estimate was above the census count (undercount) while the V2010 estimate was below the census count (overcount). For a small percentage of counties (6.3), the V2010 showed an undercount while the Subnational DA estimates showed an overcount.

We explore the geographic distribution of overlap and differences between net coverage error using the Subnational DA and Vintage 2010 estimates (Figures 3 and 3a). Figure 3 maps counties by the categories reported in Table 6: undercount in both, undercount in Subnational DA / overcount in V2010, overcount in DA / undercount in in Vintage 2010, and overcount in both. Counties showing an undercount in both set of estimates are found in all states and regions, but are most concentrated in the **South** and **Southwest**.

Counties where both series identified overcount are found primarily in New England, the Great Plains, and rural counties in the **West**. Figure 3a maps only the counties that flipped coverage patterns between the two sets of benchmark estimates. Counties where the Subnational DA shows an undercount, but V2010 shows an overcount, are geographically dispersed throughout the United States. Counties where the Subnational DA showed an overcount and Vintage 2010 showed an undercount were located in the Great Plains and Western states.

The map in Figure 2 shows that net coverage error of young children is related to spatial patterns of race and ethnic distributions in the United States. For example, we know that the border counties in the Southwest, which show undercounts of 10 percent or more, have large Hispanic populations.

Cluster Analysis Results

Since we did not produce Subnational DA estimates by race and Hispanic origin, we use cluster analysis to groups counties based on their population and housing characteristics. We then analyze the median net coverage error for each cluster. Table 7 reports the seven county clusters that we identified in our analysis and the median coverage error in the Subnational DA and V2010. The median net coverage error for all counties using the Subnational DA estimates was -2.3 percent compared to -0.2 percent in the Vintage 2010 estimates. We also map the counties by cluster (Figure 4). See Appendix B for details on how the clusters were obtained.

The clusters, in order of median net coverage error in the Subnational DA estimates are: **Majority Black/White**, **Majority American Indian/Alaska Native**, **High Hispanic proportion**, **Immigrant destinations**, **Prison and Military**, **Average characteristics**, and **Majority White**.

The **Majority Black/White** are counties that showed the highest proportions of the Non-Hispanic Black population in the 2010 Census. The remaining population is primarily Non-Hispanic White, while other race and Hispanic origin groups are underrepresented. They are located mainly in the **South**. (Figure 4). This cluster had the highest median net coverage error in the Subnational DA estimates with -5.5 percent (Table 7).

Counties in the **Majority American Indian/Alaska Native** cluster are located in Alaska, the Four Corners region, and parts of Montana, South Dakota, and North Dakota. The median net coverage error of this cluster was -5.0 percent.

The **High Hispanic proportion** cluster is located primarily in the Southwest and Western states. This cluster had a median net coverage error of -3.4 percent.

The **Immigrant destinations** cluster saw a median net coverage error of -2.5 percent. This cluster had the highest levels of immigration in the Population Estimates Program over the period 2000-2009. Counties in this group are racially diverse and include large urban centers, college counties, and scattered immigrant enclaves in rural areas. This cluster showed a high average proportion of the Non-Hispanic Asian population in Census 2010.

Counties in the **Prison and Military** cluster, with a median net coverage error of -2.3 percent, are not concentrated in any particular region or state. They are characterized by large Group Quarters populations and skew more male than all the other clusters.

The median net coverage error for the **Average characteristics** cluster is -2.0. This cluster has the largest number of counties (1,402) and is made up of counties with demographic characteristics that are similar to the national average, with a slight skew towards the White population. These counties are geographically dispersed across the United States.

The final cluster, **Majority White**, is made up of counties that have large White populations and tend to have older age distributions. These counties had a median net coverage error of -0.4 percent. The **Majority White** cluster is concentrated in New England, the Great Plains, and rural counties in the West.

Limitations

There are several limitations in our subnational study of the net undercount of young children. As noted previously, there can be discrepancies between different sources of vital events (potentially a misclassification of place of residence and occurrence detail exacerbated by complex geography). We mitigated some of these issues by supplementing the NCHS with state data. However, our method of reconciliation is simple, and there are potentially more sophisticated ways of blending these different types of data. Additionally, it is possible that we introduced more error with this choice of reconciliation.

Despite these issues, data on recent vital events are of relative high quality and coverage. (O'Hare, 2015). On the other hand, data and methods required to estimate domestic and international migration introduce more uncertainty and require simplifying assumptions. This limitation of the Subnational DA estimate is somewhat mitigated for this project because the estimates are largely driven by high quality vital records and not migration estimates for ages 0 to 4.

There are limitations inherent in the domestic migration component. First, not everyone files taxes, and there are segments of the population not covered by these data. Namely, there are income thresholds that exclude low income individuals, and differential by coverage race and Hispanic origin has been identified (Miller, 2014). This coverage issue can bias the domestic migration estimates. Additionally, IRS data and other administrative data source are not collected with the intent of producing population estimates and are susceptible to measurement error and varied levels of data quality.

Conclusion

The population age 0 to 4 was undercounted in the 2010 Census at a higher rate than any other age group. Understanding the coverage of young children in the census and surveys has been a priority for the U.S. Census Bureau in recent years (U.S. Census Bureau 2014). However, very little of this research has focused on the coverage patters at the state and county levels because the 2010 DA estimates were only produced at the national level. While the Vintage 2010 Population Estimates for young children have similarities to the DA method (not based on the 2000 Census but developed primarily using birth data), there are significant limitations to those data.

In this paper, we have produced state and county DA estimates for the population age 0 to 4 on April 1, 2010. The Subnational DA estimates are an improvement over the Vintage 2010 Population Estimates for several reasons 1) we have updated the birth records, 2) improved the method for processing vital records, 3) developed domestic migration estimates that are specific for this cohort, 4) none of the components are based on 2000 Census data, and 5) incorporated data from Mexico on U.S.-born migrants. Comparing the Subnational DA estimates to the census counts provides a more accurate estimate of the undercount of young children in the 2010 Census at the state and county levels.

Compared to the Vintage 2010 analysis, we show that using the Subnational DA estimates to calculate net coverage error produces more overall counties with an undercount for young children in the 2010 Census. At the same time, the Subnational DA estimates reduce the extreme values for both undercounts and overcounts. These differences in net coverage patterns between the Subnational DA and the Vintage 2010 series are reiterated at the state level, since we sum the counties to obtain state estimates.

For this project, we did not produce Subnational DA estimates by race and Hispanic origin. However, this is a strength and not a limitation of our analysis. We are able to see a clearer picture of the correlation between race, ethnicity, and coverage by conducting a cluster analysis of counties, because the analysis is not confounded by the difficulty in assigning race and Hispanic origin to our input data.[6]

[6] Prior unpublished research by the authors comparing the Vintage 2010 Population Estimates to the 2010 Census showed large differences by race and Hispanic origin in some counties that may indicate classification error and not coverage issues

The results show that the net median net coverage error for counties varies by different clusters that we identified using population and housing data. Counties in the **Majority Black/White** cluster, which is geographically concentrated in the **South**, had the largest median net coverage error (-5.5 percent). This is considerably larger than the **Majority White** cluster which had the lowest median net coverage error (-0.4 percent) and are geographically located in New England, the Great Plains, and rural counties in the **West**. Again, the cluster analysis allows us to see the overlap between race, ethnicity, and coverage for young children.

In future research, this work can be expanded to include general hard-to-count attributes identified by other studies at the national level – linguistic isolation, poverty, complex living arrangements – with our subnational results for young children (O'Hare 2014, U.S. Census Bureau 2016b). Specifically, examining the "where" (identifying states and counties) and the "why" (determinants) of the undercount of young children in the 2010 Census could improve census operations in the future by directing resources.

In 2020, DA will again be used to measure coverage in the decennial census. As we prepare for the 2020 Census, it is important that we explore methods that can evaluate the quality of the census. While DA has successfully been used to measure coverage by age, sex, race, and Hispanic origin, it has not been used to produce subnational measures of coverage. This research shows how the methods of DA could be used to measure net coverage error at the state and county levels for some cohorts.

References

Jensen, Eric, B. Megan Benetsky. Anthony Knapp. (2018). "A Sensitivity Analysis of the Net Undercount for Young Hispanic Children in the 2010 Census." Poster Presented at the 2018 annual meetings of the Population Association of America, Denver, Colorado.

Layne, Mary. Deborah Wagner. Cynthia Rothhaas. (2014). Estimating Record Linkage False Match Rate for the Person Identification Validation System. Working Paper Number: CARRA-WP-2014-02. https://census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-02.pdf

Miller, Esther M. Joseph Bowman. "Comparing IRS Exemptions to 2010 Census Population Counts." Presented at the Applied Demography Conference San Antonio, Texas January 8-10, 2014. http://demographics.texas.gov/Resources/Presentations/ADC/2014/ADC2014_3D_Miller.pdf

O'Hare, William. (2015). The Undercount of Young Children in the U.S. Decennial Census. Springer Briefs in Population Studies.

U.S. Census Bureau (2017a). Investigating the 2010 Undercount of Young Children – Examining Data Collected during Coverage Follow-up. 2020 Census Program Memorandum Series. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/final-analysis/2020-2017_05-undercount-children-examining-data.html.

U.S. Census Bureau (2017b). Investigating the 2010 Undercount of Young Children – Child Undercount Probes. 2020 Census Program Memorandum Series. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/final-analysis/2020-2017_03-undercount-children-probes.html.

U.S. Census Bureau (2017c). Investigating the 2010 Undercount of Young Children – Analysis of Coverage Measurement Results. 2020 Census Program Memorandum Series. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/final-analysis/2020-2017_04-undercount-children-analysis-coverage.html.

U.S. Census Bureau (2014). The Undercount of Young Children. U.S. Census Bureau, Washington DC (February 2014). https://www.census.gov/content/dam/Census/library/working-papers/2014/demo/2014-undercount-children.pdf

Tables and Figures

Table 1a. Inventory of Data Sources

Source	***Description***	***Year(s)***
BIRTHS		
NCHS (National Center for Health Statistics)	Final person level data on all live births by maternal characteristics and residence and characteristics of births	2005-2010
FSCPE (Federal-State Cooperative for Population Estimates	Counts of live births by residence, and calendar year by county supplied by FSCPE members (state demographers, academics, etc.) to the Census Bureau	2008-2010
State Public Data	Counts of live births by maternal characteristics and characteristics of births published on state vital statistics or public health office websites	2005-2007[7]
DEATHS		
NCHS	Final person level data on all deaths by decedent characteristics and county of residence	2005-2010
NET INTERNATIONAL MIGRATION		
ACS /PRCS (American Community Survey/Puerto Rico Community Survey)	Annual survey on demographic, social, economic characteristics of persons and households that replaces the long form questionnaire of the decennial census; used 5-year survey estimates on residence 1-year ago and year of entry	2006-2010
2010 Mexican Census	Mexican Census – Short questionnaire on demographics and a sample from long questionnaire on migration and nativity	2010
ENOE (National Survey of Occupation and Employment	Monthly labor force survey conducted in Mexico by the National Institute of Statistics and Geography (INEGI)	2010
NCHS	Final data on all live births by maternal characteristics and residence and characteristics of births; used maternal place of birth to distribute net native migration component	2010
Other Foreign Censuses/ Population Registers	Population census and register data from other counties are used to develop estimates of net native international migration	1990 and 2000

[7] We use NCHS data for Nevada, Louisiana, Montana, Maine, Minnesota, and Vermont for some years.

Table 1b. Inventory of Data Sources

Source	***Description***	***Year(s)***
NET DOMESTIC MIIGRATION		
SSA NUMIDENT (Social Security Administration)	Person level data on all Social Security Numbers ever assigned; includes date of birth, sex, and place of birth	2017 (most updated version)
IRS 1040 Tax Return Data (Internal Revenue Service)	Person level data on all individuals who appear on a tax return as the tax filer, spouse, or dependent (contains address information)	Tax years 2004-2009 (all returns filed 2005-2010)
CLUSTER ANALYSIS		
2010 Census DP-1	Demographic and housing characteristics of all counties (percent Non-Hispanic, Median age, Percent Occupancy, etc.)	2010
Vintage 2010 Population Estimates (without Challenges and Special Census Results)	Estimates of county population by age group, sex, race, and Hispanic origin 2000-2010; used estimates of cumulative net international migration by county over 2000-2009 and over 2008-2009	2000-2009, 2008-2009

Table 2. Relationship between Subnational DA and Vintage 2010 Net Coverage Errors of Young Children in the 2010 Census by State

Coverage Pattern	**Number of States**	**Percent of States**
Undercount in both	45	90.0
Undercount in DA / Overcount in V2010	4	8.0
Overcount in DA / Undercount in V2010	1	2.0
Overcount in both	0	0.0
Total	50	100

Source: 2010 Census, Subnational DA Estimates, and Vintage 2010 Population Estimates.

Table 3. State Level Net Coverage Errors of Young Children in the 2010 Census: Subnational DA Estimates and Vintage 2010 Population Estimates

State	Subnational DA	Vintage 2010	Census 2010	Net Coverage Error		Rank (1=Highest Net Error)		
				DA	V10	DA	V10	Change in Rank
Florida	1,144,813	1,160,358	1,073,506	-6.2	-7.5	1	3	-2
Mississippi	224,723	220,930	210,956	-6.1	-4.5	2	13	-11
Delaware	59,260	59,053	55,886	-5.7	-5.4	3	9	-6
Georgia	727,401	739,429	686,785	-5.6	-7.1	4	5	-1
Virginia	538,688	531,833	509,625	-5.4	-4.2	5	15	-10
Alabama	322,275	317,230	304,957	-5.4	-3.9	6	17	-11
New Jersey	570,719	554,575	541,020	-5.2	-2.4	7	32	-25
Texas	2,030,879	2,079,561	1,928,473	-5.0	-7.3	8	4	4
Maryland	383,502	380,425	364,488	-5.0	-4.2	9	14	-5
West Virginia	109,168	106,852	104,060	-4.7	-2.6	10	29	-19
Hawaii	91,693	90,611	87,407	-4.7	-3.5	11	21	-10
New York	1,210,688	1,220,150	1,155,822	-4.5	-5.3	12	10	2
Tennessee	427,154	422,263	407,813	-4.5	-3.4	13	22	-9
Massachusetts	384,311	385,403	367,087	-4.5	-4.8	14	12	2
Rhode Island	60,041	59,443	57,448	-4.3	-3.4	15	24	-9
Kentucky	294,935	289,308	282,367	-4.3	-2.4	16	33	-17
Arizona	475,797	506,523	455,715	-4.2	-10.0	17	1	16
Illinois	871,483	885,570	835,577	-4.1	-5.6	18	7	11
California	2,633,187	2,736,963	2,531,333	-3.9	-7.5	19	2	17
Louisiana	326,349	318,954	314,260	-3.7	-1.5	20	38	-18
Oklahoma	274,152	274,442	264,126	-3.7	-3.8	21	19	2
South Carolina	313,618	312,916	302,297	-3.6	-3.4	22	23	-1
Missouri	404,633	400,241	390,237	-3.6	-2.5	23	31	-8

Source: 2010 Census, Subnational DA Estimates, and Vintage 2010 Population Estimates.

Legend

🟨	"Harder to count" in Subnational DA: Moved ***up*** in rank by 10 or more
🟩	"Easier to count" in Subnational DA: Moved ***down*** in rank by 10 or more

Table 3. State Level Net Coverage Errors of Young Children in the 2010 Census: Subnational DA Estimates and Vintage 2010 Population Estimates (cont'd)

State	Subnational DA	Vintage 2010	Census 2010	Net coverage Error		Rank (1=Highest Net Error)		
				DA	V10	DA	V10	Change in Rank
Pennsylvania	754,735	748,583	729,538	-3.3	-2.5	24	30	-6
North Carolina	653,195	656,138	632,040	-3.2	-3.7	25	20	5
Arkansas	204,254	204,042	197,689	-3.2	-3.1	26	26	0
Nebraska	135,940	134,368	131,908	-3.0	-1.8	27	37	-10
Ohio	742,570	736,309	720,856	-2.9	-2.1	28	35	-7
Connecticut	208,112	208,654	202,106	-2.9	-3.1	29	25	4
New Mexico	149,140	153,260	144,981	-2.8	-5.4	30	8	22
Indiana	446,415	444,273	434,075	-2.8	-2.3	31	34	-3
New Hampshire	71,685	71,869	69,806	-2.6	-2.9	32	28	4
Minnesota	364,254	362,152	355,504	-2.4	-1.8	33	36	-3
Iowa	206,849	203,533	202,123	-2.3	-0.7	34	43	-9
South Dakota	60,925	59,851	59,621	-2.1	-0.4	35	45	-10
Kansas	209,957	207,396	205,492	-2.1	-0.9	36	41	-5
Washington	448,624	457,061	439,657	-2.0	-3.8	37	18	19
Colorado	350,508	361,483	343,960	-1.9	-4.8	38	11	27
Michigan	606,458	602,137	596,286	-1.7	-1.0	39	40	-1
Alaska	54,899	54,751	53,996	-1.6	-1.4	40	39	1
Nevada	190,376	200,684	187,478	-1.5	-6.6	41	6	35
Wisconsin	363,924	360,951	358,443	-1.5	-0.7	42	42	0
North Dakota	45,029	43,644	44,595	-1.0	2.2	43	50	-7
Wyoming	40,474	40,062	40,203	-0.7	0.4	44	47	-3
Vermont	32,155	31,675	31,952	-0.6	0.9	45	49	-4
Oregon	238,922	247,844	237,556	-0.6	-4.2	46	16	30
Maine	69,890	69,683	69,520	-0.5	-0.2	47	46	1
Utah	265,193	272,392	263,924	-0.5	-3.1	48	27	21
Montana	62,675	62,056	62,423	-0.4	0.6	49	48	1
Idaho	121,264	122,530	121,772	0.4	-0.6	50	44	6

Source: 2010 Census, Subnational DA Estimates, and Vintage 2010 Population Estimates.

Legend

🟨	"Harder to count" in Subnational DA: Moved ***up*** in rank by 10 or more
🟩	"Easier to count" in Subnational DA: Moved ***down*** in rank by 10 or more

Table 4. Summary of County Level Net Coverage Error of Young Children in the 2010 Census using Subnational DA and Vintage 2010 Estimates

Net Coverage Category	Subnational DA		Vintage 2010	
	Number	Percent	Number	Percent
10 or more percent undercount	203	6.5	279	8.9
1 to less than 10 percent undercount	1,734	55.2	1,169	37.2
Full coverage (within +/- 1 percent)	471	15.0	377	12.0
1 to less than 10 percent overcount	611	19.4	922	29.3
10 or more percent overcount	124	4.0	396	12.6
Total	3,143	100.0	3,143	100.0

Source: 2010 Census, Subnational DA Estimates, and Vintage 2010 Population Estimates.

Table 5. Mean Absolute Percent Error (MAPE) and Mean Algebraic Percent Error (MALPE) of Net Coverage of Young Children in the 2010 Census by County Population Size

Total County					
Population	Number of	Subnational DA		Vintage 2010	
In 2010 Census	Counties	MAPE	MALPE	MAPE	MALPE
Less than 5,000	303	9.6	3.1	23.0	13.7
5,000 to 9,999	395	6.2	-0.3	10.3	3.6
10,000 to 19,999	607	4.9	-1.7	6.8	1.1
20,000 to 64,999	1,033	4.0	-2.5	4.8	-0.6
65,000 to 99,999	227	3.6	-2.7	3.9	-0.7
100,000 to 249,999	317	3.4	-3.1	3.4	-1.8
250,000 to 499,999	133	3.9	-3.6	4.0	-3.5
500,000 or more	128	4.3	-4.2	6.4	-6.1
All Counties	3,143	4.9	-1.7	7.4	1.2

Source: 2010 Census, Subnational DA Estimates, and Vintage 2010 Population Estimates.

Table 6. Relationship between Subnational DA and Vintage 2010 Net Coverage Errors of Young Children in the 2010 Census by County

Coverage Pattern	Number of Counties	Percent of Counties
Undercount in both	1,421	45.2
Undercount in DA / Overcount in V2010	772	24.6
Overcount in DA / Undercount in V2010	198	6.3
Overcount in both	752	23.9
Total	3,143	100.0

Source: 2010 Census, Subnational DA Estimates, and Vintage 2010 Population Estimates.

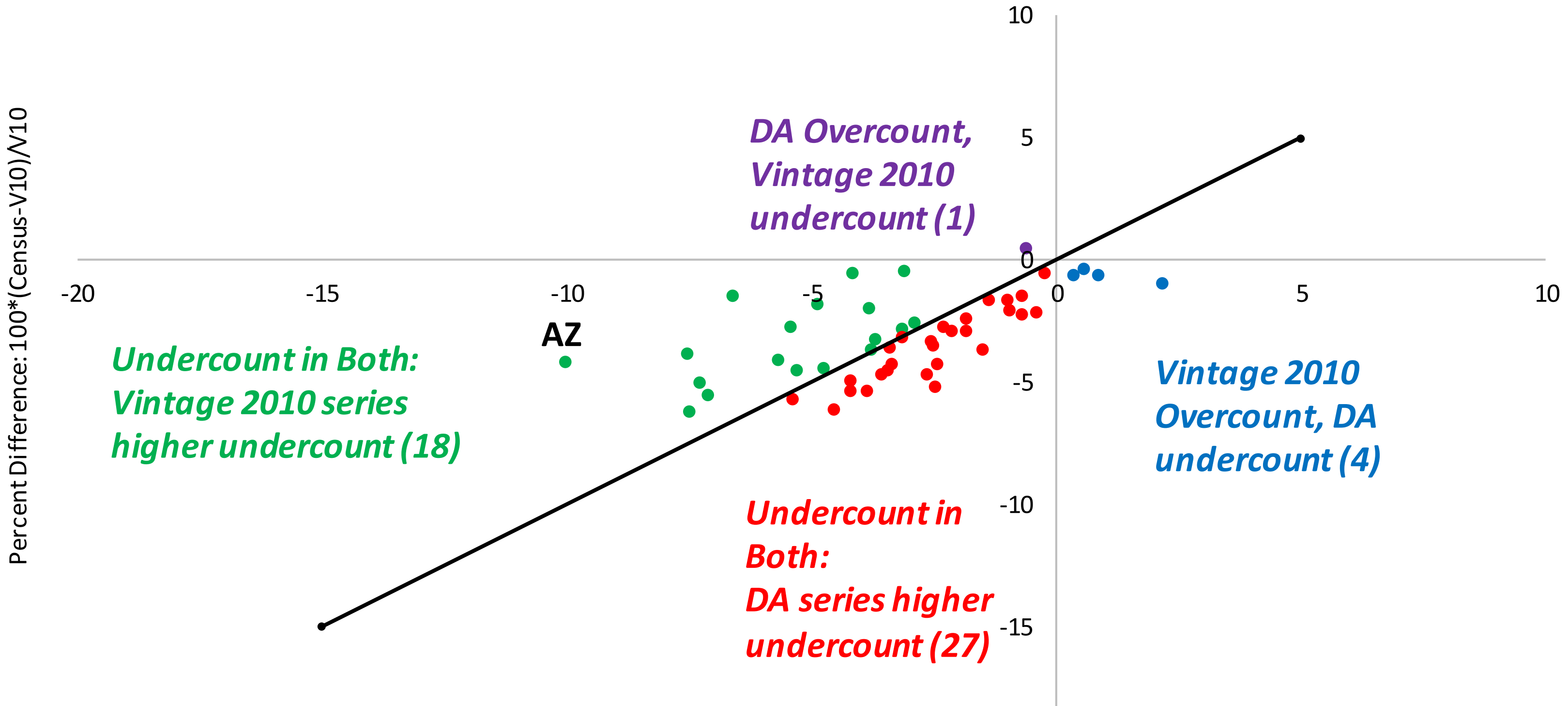
Table 7. Median Net Coverage Error of Young Children by County Clusters formed from Demographic and Housing Characteristics

		Median Net Coverage Error	
County Cluster	Number of Counties	Subnational DA	Vintage 2010
Majority Black/White	393	-5.5	-5.0
Majority American Indian/Alaska Native	24	-5.0	-6.6
High Hispanic proportion	327	-3.4	-2.4
Immigrant Destinations	136	-2.5	-4.4
Prison and Military	123	-2.3	0.7
Average Characteristics	1,402	-2.0	0.3
Majority White	704	-0.4	4.3
Total	3,143	-2.3	-0.2

Source: Coverage - 2010 Census, Subnational DA Estimates, and Vintage 2010 Population Estimates.
Clusters - 2010 Census DP-1 and Vintage 2010 estimates of international migration.

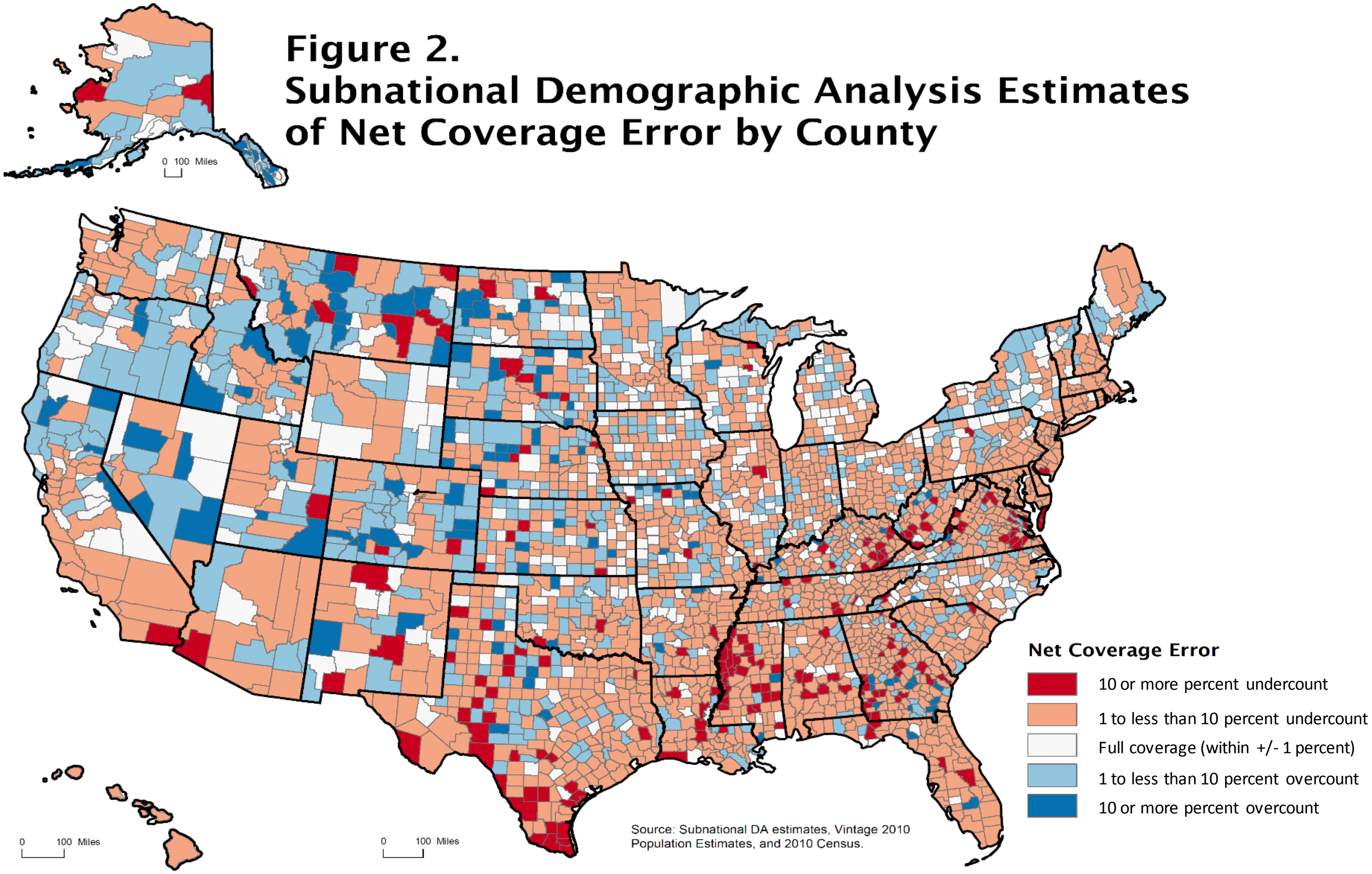
Figure 1. Net Coverage Error of Young Children in States, Percent Difference from Two Series:

Vintage 2010 Population Estimates (horizontal), Subnational DA (vertical)



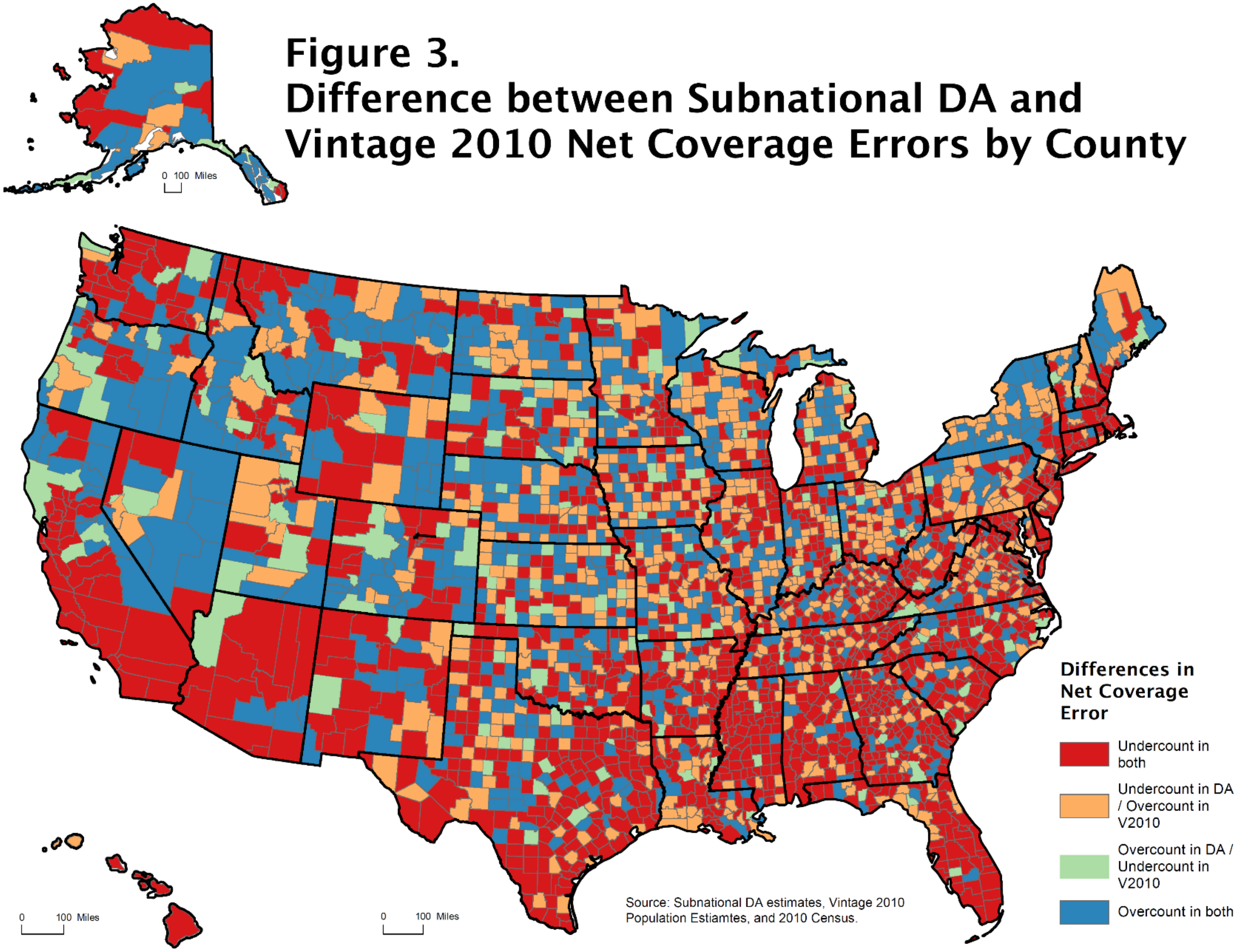
Source: Subnational DA Estimates, Vintage 2010 Popuation Estimates, and Census 2010.

Figure 2. Subnational Demographic Analysis Estimates of Net Coverage Error by County


0 100 Miles
0 100 Miles
0 100 Miles
Net Coverage Error
10 or more percent undercount
1 to less than 10 percent undercount
Full coverage (within +/- 1 percent)
1 to less than 10 percent overcount
10 or more percent overcount

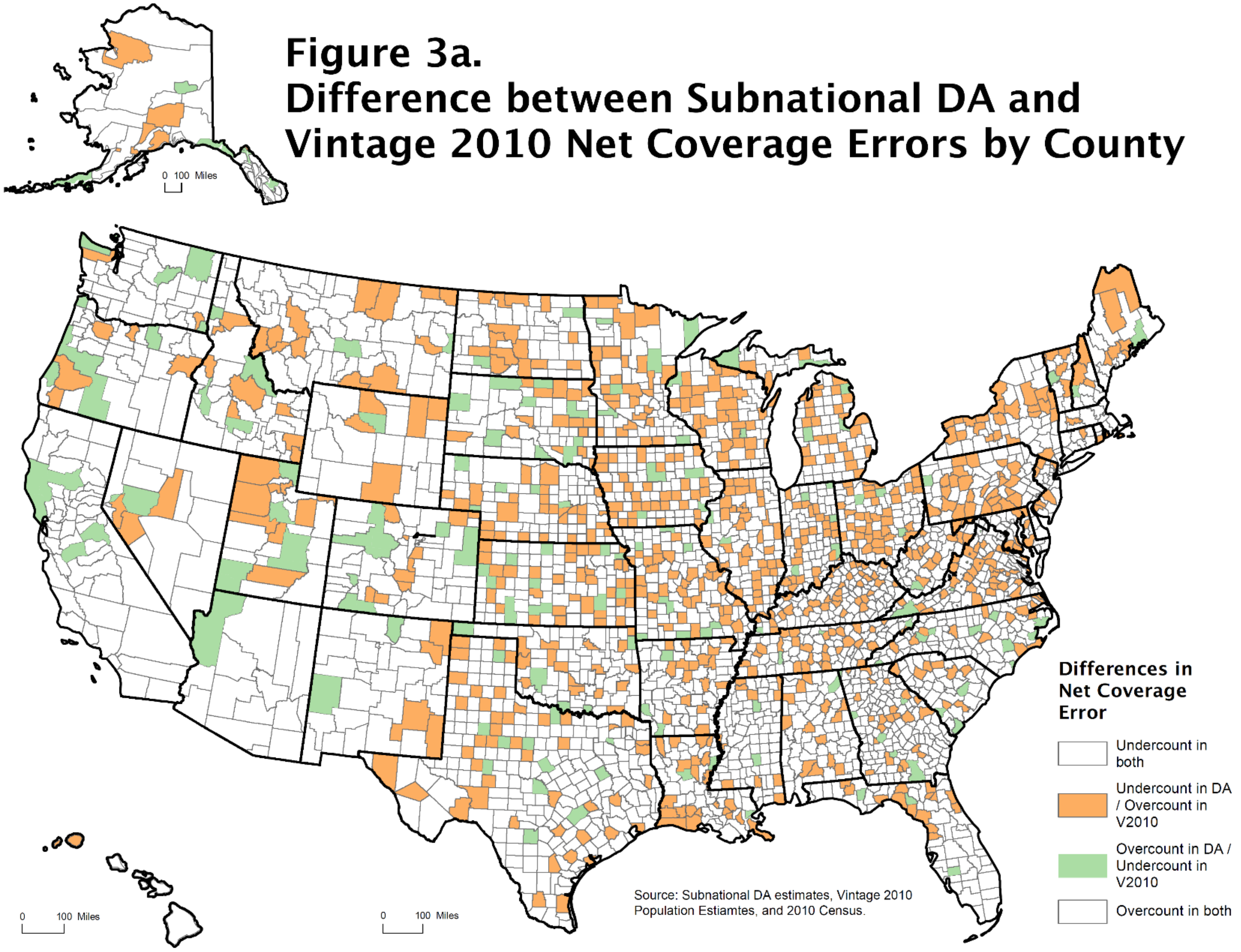
Source: Subnational DA estimates, Vintage 2010 Population Estimates, and 2010 Census.

Figure 3.
Difference between Subnational DA and Vintage 2010 Net Coverage Errors by County


0 100 Miles
0 100 Miles
0 100 Miles
Differences in Net Coverage Error
🟥 Undercount in both
🟧 Undercount in DA / Overcount in V2010
🟩 Overcount in DA / Undercount in V2010
🟦 Overcount in both

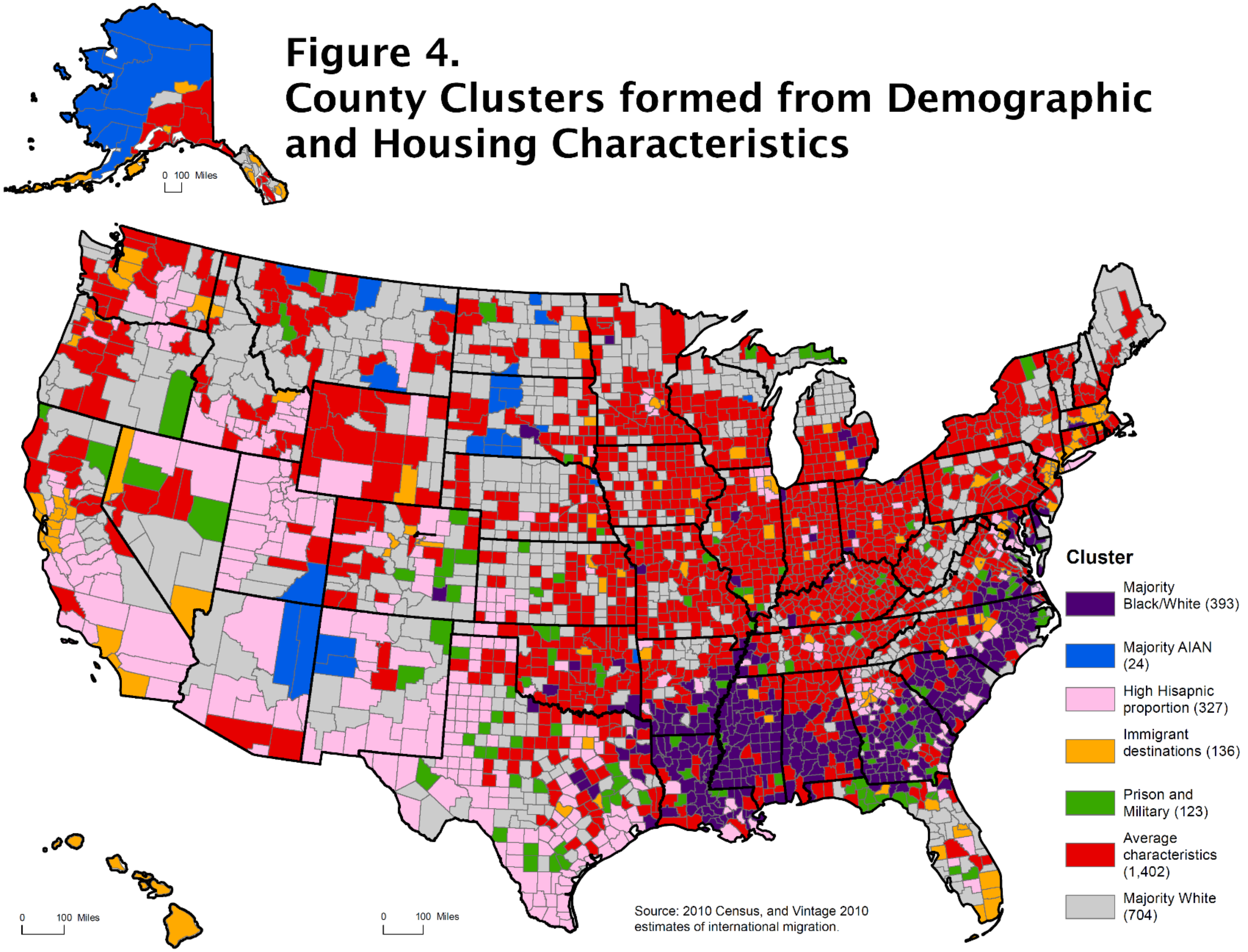
Source: Subnational DA estimates, Vintage 2010 Population Estiamtes, and 2010 Census.

Figure 3a.
Difference between Subnational DA and Vintage 2010 Net Coverage Errors by County


0 100 Miles
0 100 Miles
0 100 Miles
Differences in Net Coverage Error
□ Undercount in both
🟧 Undercount in DA / Overcount in V2010
🟩 Overcount in DA / Undercount in V2010
□ Overcount in both

Source: Subnational DA estimates, Vintage 2010 Population Estiamtes, and 2010 Census.

Figure 4.
County Clusters formed from Demographic and Housing Characteristics


0 100 Miles
0 100 Miles
0 100 Miles
Cluster
Majority Black/White (393)
Majority AIAN (24)
High Hisapnic proportion (327)
Immigrant destinations (136)
Prison and Military (123)
Average characteristics (1,402)
Majority White (704)

Source: 2010 Census, and Vintage 2010 estimates of international migration.

APPENDIX A: NCHS and State Data Reconciliation

This section describes how we combine three sources of birth data to estimate a series that preserves county distribution from local data and maintains state totals from the National Center for Health Statistics (NCHS).

In the Subnational DA series, we follow these steps to produce a county level births series for 2005-2010 that combines all 3 sources of birth data:

1. use state public county data for years 2005-2007
2. use internal FSCPE county data for years 2008-2010
3. control full series of county births to the NCHS state total

STEP 1: Combine the State Data

After appending the FSCPE data (2008-2010) to the state public data (2005-2007) to create a full time series of county birth data, we compared the series to the NCHS tallies of county births over the same time period. Below are the findings.

On average, the numeric differences between the two series remained between +/- 5 births in every year and within +/- 0.5 percent. Around 90 percent of counties showed numeric differences with the NCHS series within +/-30 births for all years and within +/- 6 percent. The states with the largest discrepancies between NCHS and state data fall within expectations (i.e. the same states noted in annual estimates production).

STEP 2: Reconcile the State Data with NCHS Data

After we create a state series of county annual births from external public health data and our internal FSPCE data and for years 2005-2010, we adjust the NCHS births to sum to the county totals from the state sources. To this end, we apply a rake factor to the NCHS births by county, year, and month as follows:

1. NCHS county births by year and month sum to county totals from state sources

   Rake factor 1: $\frac{county\ total\ from\ state\ sources}{NCHS\ county\ total}$

The next step of reconciliation preserves the NCHS state total and FSCPE county distribution by applying a rake factor that proportionally adjusts the county totals from the previous step.

2. NCHS/FSCPE adjusted county totals from previous step controlled to NCHS state totals by period

   Rake factor 2: $\frac{NCHS\ state\ total}{Adjusted\ NCHS\ state\ total}$

APPENDIX B: Cluster Analysis Details

This section explains how we conducted the cluster analysis to create groups of demographically similar counties.

We used 18 variables that describe the age structure, sex composition, race and Hispanic origin distribution, nativity and levels of immigration, and housing characteristics for each county. With the exception of the immigration indicators, all data were obtained from publicly available sources from Census 2010. The foreign born indicators were taken from the Population Estimates Program's estimates of net international migration over the periods 2000-2007 and 2008-2009.

We performed the cluster analysis in two steps:

1. Conduct Principal Component Analysis (PCA) on the demographic/housing variables
2. Cluster the output from PCA using k-means clustering to identify well-separated groups of counties with similar population and housing characteristics

The PCA was performed first to combine related demographic and housing indicators into a set of uncorrelated factors that sufficiently describe the data. Six principal components were selected by taking all eigenvalues of at least 1, examining the scree pl ot, and evaluating the variance added with each additional principal component. The first six components explain about 85 percent of the variance in the data.

The magnitude of correlation between the six principal components and the original variable was used to pair each principal component to the demographic/housing characteristic that it primarily describes:

1. Age
2. Household characteristics
3. Levels of foreign-born immigration
4. Sex
5. Hispanic origin
6. Race

The output of the PCA was used as input to the cluster analysis. The number of clusters k was identified by maximizing the cubic clustering criterion.