December 2021

METHODOLOGY FOR THE UNITED STATES POPULATION ESTIMATES: VINTAGE 2021

Nation, States, Counties, and Puerto Rico – April 1, 2020 to July 1, 2021

Populations can change in three ways: people may be born (births), they may die (deaths), or they may move (domestic and international migration). The U.S. Census Bureau's Population Estimates Program measures this change and adds it to a base population to produce updated estimates every year.

OVERVIEW

Each year, the United States Census Bureau produces and publishes estimates of the population for the nation, states, counties, state/county equivalents, and Puerto Rico.[1] We estimate the resident population for each year since the most recent decennial census by using measures of population change. The resident population includes all people currently residing in the United States.

With each annual release of population estimates, the Population Estimates Program revises and updates the entire time series of estimates from April 1, 2020 to July 1 of the current year, which we refer to as the vintage year. We use the term "vintage" to denote an entire time series created with a consistent population starting point and methodology. The release of a new vintage of estimates supersedes any previous series and incorporates the most up-to-date input data and methodological improvements.

The population estimates are used for federal funding allocations, as controls for major surveys including the Current Population Survey and the American Community Survey, for community development, to aid business planning, and as denominators for statistical rates. Overall, our estimates time series from 2000 to 2010 was very accurate, even accounting for ten years of population change. The average absolute difference between the final total resident population estimates and 2010 Census counts was only about 3.1 percent across all counties.[2]

We produce estimates using a cohort-component method, which is derived from the demographic balancing equation:

Population Base + **Births** − **Deaths** + **Migration** = **Population Estimate**

The population estimate at any given time point starts with a population base (e.g. the last decennial census or the previous point in the time series), adds births, subtracts deaths, and adds net migration (both international and domestic).[3] The individual methods we use account for additional factors such as input data availability and the requirement that all estimates be consistent by geography and age, sex, race, and Hispanic origin.

This document describes the input data, methodology, and processes for the creation of population estimates for the nation, states, counties, state/county equivalents, and Puerto Rico. We begin with a short discussion on consistency in the estimates, describe the input data, and detail the processes by which we produce estimates.

[1] The methodologies for developing population estimates for incorporated places and minor civil divisions (cities and towns) and housing unit estimates are covered in separate documents.

[2] For more information on the accuracy of the population estimates, see https://www.census.gov/library/working-papers/2013/demo/POP-twps0100.html .

[3] Domestic migration sums to 0 at the national level and therefore has no effect on the estimates.

Estimates Consistency, Controlling, and the Residual

We produce the estimates using a “top-down” approach. Given that it is generally more reliable to estimate the change of a larger population, we begin by estimating the monthly population at the national level by age, sex, race, and Hispanic origin. We then produce estimates of the total annual populations of counties, which we sum to the state level. With the national characteristics, state total, and county total estimates created, we produce estimates of states and counties by age, race, sex, and Hispanic origin.

One of our key estimates principles is that all of the estimates we produce must be consistent across geography and demographic characteristics. For example, the sum of the county total populations must equal the total national population, and the sum of a particular race group within a state’s counties must equal the total of that particular race group in the state. Since our various estimates products and processes use slightly different input data and methodology, they often do not generate this consistency automatically. Consequently, we adjust the final estimates to be consistent. As a result, the demographic components of change do not account for all of the year-to-year change in the estimates series. We call the difference between the result of the balancing equation and the final estimate the residual.

The national population estimates by characteristics do not contain a residual. This is because they are made first and are not required to sum to any pre-defined total. The balancing equations for the subnational processes initially produce what we call “uncontrolled” estimates. In order to ensure consistency, we use a process called controlling or raking. This involves calculating a rake factor as the control total (to which data must sum) divided by the sum of the numbers we wish to control (the initial estimated values).

$$Rake = \left(\frac{Control\ Total}{\sum(Uncontrolled\ Values)}\right)$$

We multiply this rake factor by the uncontrolled values to generate “controlled” estimates. In the simple case where the goal is to sum to a column total, this is fairly straightforward. However, deriving state and county population estimates by characteristics requires a slightly more complicated process. Since we produce national estimates by characteristics and state/county totals first, state and county characteristics need to use a two-way raking system. For example, state characteristics are required to be consistent with national characteristics and state total estimates (see the section on state and county characteristics).

The controlling process usually produces estimates that sum to a predefined total but are not integers. Because we require estimates in integer form, we round these data to remove the decimal values. Applying a simple rounding algorithm may upset the consistency established in the controlling process. To account for this, we use a variety of controlled rounding procedures (e.g., greatest mantissa or two-way controlled rounding).

Base Population

The population estimates base is the starting point for each vintage of population estimates. Over recent decades, the decennial census typically provided all the necessary detail for the estimates base. However, the 2020 Census could not be similarly adopted for this purpose due to several challenges.

First, the disclosure avoidance system applied to the 2020 Census counts had an impact on what variables would be available in the official (i.e. protected via differential privacy) data. This included several variables required for estimates processing, such as "modified race"[4] (race variable featuring redistributed "Some other race" responses into the race groups defined by the Office of Management and Budget in 1997), the Master Address File ID (used to implement annual boundary updates), and variables necessary for data record linkages with administrative records (used to assign demographic characteristics for births and domestic migration).

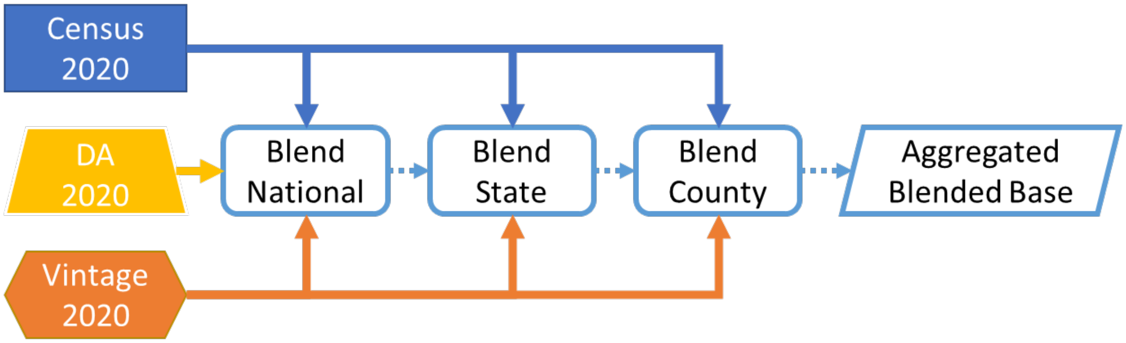
Second, the COVID-19 pandemic introduced significant delays to both enumeration and data processing schedules. At the time of Vintage 2021 estimates production, official decennial data by the full age, sex, race, Hispanic origin, and universe (e.g., household population) detail required for processing were not available.

Third, because of these schedule delays, the Population Estimates Program has not yet completed its evaluation of the 2020 Census data to determine its suitability for the specific use case of a full-detail estimates base population.

Due to these challenges, the Population Estimates Program developed a process for integrating three data sources at varying levels of detail to produce what we refer to as the Blended Base. The Blended Base represents the most detail from alternate sources we could confidently incorporate into the estimates base with the time that was available.

- **2020 Census PL 94-171 Redistricting File**: Nation, state, county, and Puerto Rico total population counts
- **2020 Demographic Analysis (DA)**[5] **Estimates**: National population estimates by age and sex
- **Vintage 2020 Postcensal Population Estimates**: Nation, state, and county population estimates by age, sex, race, Hispanic origin, and population universe; and Puerto Rico Commonwealth and municipio population estimates by age, sex, and population universe

Figure 1. Blended Base Process for the Nation, States, and Counties



As depicted in Figure 1, the Blended Base process uses a top-down methodology which is very similar to how the postcensal population estimates are developed every year. We create blended national-level data by first applying the 2020 DA national population distribution by single year of age and sex to the 2020 Census totals. We then rake the full-detail Vintage 2020 estimates to the combined DA and Census data, resulting in a dataset that integrates the 2020 Census, 2020 DA, and the Vintage 2020 estimates. At the national level, then, it is accurate to say that

[4] In our estimates processing, we modify the Census race categories to be consistent with the race categories that appear in our input data. To learn more about the "Modified Race" process, go to http://www.census.gov/programs-surveys/popest/technical-documentation/research/modified-race-data.html.

[5] The 2020 DA estimates of the national population by age, sex, race, and Hispanic origin on April 1, 2020 are developed from current and historical vital records, estimates of international migration, and Medicare records. The DA estimates are independent from the 2020 Census and are used to calculate net coverage error, one of the two main ways the U.S. Census Bureau uses population estimates to measure coverage of the census. For more information, see https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/da.html.

population totals come from the decennial census, age and sex detail comes from DA, and race and Hispanic origin detail comes from the Vintage 2020 estimates. Using the DA data allows the Blended Base to make some adjustments for some known limitations in past decennial censuses, such as the undercoverage of young children.

We then rake the Vintage 2020 state-level estimates to the national level Blended Base by full detail and to the 2020 Census state totals. This allows us to retain the benefits of the national Blended Base while keeping the final populations consistent with previously released 2020 Census data. We develop the county-level Blended Base data using the same method, raking the Vintage 2020 county estimates, in Vintage 2020 geographic boundaries, to the state Blended Base and the 2020 Census county total counts. Finally, we round, aggregate the county-level estimates to ensure geographic consistency, and model additional detail required for our estimates processing (e.g., population universes or quarter years of age) using the Vintage 2020 data.

The development of the Blended Base for Puerto Rico follows the same steps. The main differences are that there is no DA control available for Puerto Rico and that the annual Puerto Rico estimates are only produced by age, sex, and population universe. The Puerto Rico Commonwealth Blended Base is developed by raking the Vintage 2020 April 1, 2020 population by age and sex directly to the 2020 Census total counts. Municipio data then follow the same process as U.S. counties, being raked to both the Puerto Rico Commonwealth Blended Base and the municipio 2020 Census total counts.

Group Quarters

We estimate the group quarters (GQ) population every year by single year of age, sex, race, Hispanic origin, and facility type.[6] The GQ method begins with an estimates base derived from the previous decennial census. We assume that the population in GQ remains constant throughout the decade unless we receive updated data on GQ population change.

Information on change to the base GQ population comes from our annual Group Quarters Report (GQR). The GQR consists of time series data from the branches of the military, the Department of Veterans Affairs, and our state partners in the Federal-State Cooperative for Population Estimates (FSCPE). Our data providers supply data at the facility level, which allows us to aggregate to all the other estimates geographies (e.g., counties and states). We use the submitted data to calculate a year-to-year change, which we then apply to the GQ population in the estimates base.

Once we have a times series of total GQ population at the facility level, we aggregate the facility-level data to the national level and apply the 2010 Census distribution of age, sex, race, and Hispanic origin detail by major facility type to generate estimates of the GQ population by demographic characteristics. We also apply the county distribution of age, sex, race, and Hispanic origin to the county level totals. To ensure consistency, we control the county characteristics to the national characteristics and the subcounty totals to the new county totals. Finally, we aggregate the data to the necessary levels for estimates production (e.g., three age groups for county totals production and full demographic detail for state characteristics production).

Vital Statistics

Vital statistics encompass two of the core components of the demographic equation: births and deaths. We receive data on vital statistics from the National Center for Health Statistics (NCHS) and the FSCPE. NCHS data are derived from birth and death certificates across the United States. Births data include date of birth, sex of child,

[6] The seven major GQ facility types utilized in estimates production are: correctional institutions, juvenile institutions, nursing homes, other institutional facilities, college dormitories, military housing, and other noninstitutional facilities. While we do not release data on GQ by facility type, we do use them to calculate population universes such as “civilian noninstitutionalized.”

residence and age of mother, and race and Hispanic origin of both mother and father. Deaths data include residence, age, sex, race, and Hispanic origin of each decedent, and the date each death occurred. The FSCPE contributes data on the geographic distribution of recent vital events within their respective states. Vital events data in the population estimates also include the results of our own short-term projections.

In general, the births and deaths data we receive from NCHS have a two-year lag. This means that the most recent final data we have on births and deaths by geographic and demographic detail for each vintage of estimates refer to the calendar year two years prior to the vintage year. For example, the most current full-detail births and deaths data we used in Vintage 2021 were from calendar year 2019. Additionally, for Vintage 2021 we had NCHS monthly provisional total numbers of births and deaths at the national level for all months of 2020. To account for changes to natality resulting from the COVID-19 pandemic, we also incorporated monthly total births for the nation in the first quarter of 2021 and used recent trends to project births for the second quarter of the year. To reflect the impact of COVID-19 on deaths, we had data for the first half of 2021 that includes recent trends and patterns of excess mortality from the pandemic. Essentially, the NCHS data are used in conjunction with the data received from the FSCPE to create short-term projections that approximate the final NCHS data by characteristics.

We also modify the NCHS births and deaths data to comply with our process. The births data require three changes. Since 2016, all 50 states and the District of Columbia have reported parents' race data to NCHS in the 1997 OMB race categories (non-Hispanic single-race White, non-Hispanic single-race Black or African American, non-Hispanic single-race American Indian and Alaska Native, non-Hispanic single-race Asian, non-Hispanic single-race Native Hawaiian and Other Pacific Islander, and Hispanic). NCHS also provides race data in the 1977 OMB race categories (White; Black; American Indian, Eskimo or Aleut; and Asian or Pacific Islander) where parents' race data are only classified into one race group. For our purposes, we first convert the race data from the 1977 standards into the newer 1997 classification utilizing a race bridging method designed by NCHS and the United States Census Bureau to make the multiple-race and single-race data comparable.[7]

Second, as birth certificates include only data on the race and Hispanic origin of the parents, not the child, we impute the race of the child through our "Kidlink" process.[8] This approach uses the combined distributions of mothers', fathers', and children's race and Hispanic origin from the 2010 Census to impute children's race and Hispanic origin.

Third, we adjust for inconsistencies between the imputed race and Hispanic origin distributions of births compared to the base population under age 1 in the 2010 Census. This benchmarking process allows us to adjust the overall race and Hispanic origin distribution of births to create a "census-consistent" time series of births.

We also make modifications to the NCHS deaths data. Although we often have direct information on the race and Hispanic origin of the decedent, deaths are still coded in many states according to the 1977 OMB race categories. We use the same race bridging process for deaths that we use to convert births into the 1997 race and Hispanic origin categories used in estimates production.

While we make no additional adjustments to deaths occurring to people under 70 years of age, we do modify death records for persons age 70 or over. Reporting of age at older ages is generally less reliable than at younger ages.[9] To address this issue, we redistribute all deaths occurring to the aggregate population 70 years and older by sex, race, and Hispanic origin to single year of age (70 to 99 and 100+ years) using life-table-based death rates.[10]

[7] For more information on the NCHS race-bridging factors, see http://www.cdc.gov/nchs/nvss/bridged_race.htm.

[8] For more information on the Kidlink process, see https://nces.ed.gov/FCSM/pdf/Guarneri_2012FCSM_X-B.pdf.

[9] For more information on age reporting at older ages, see http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_07.pdf.

[10] To derive the death rates for the age-70-and-older population, we employ life tables based on annual 2000-2010 NCHS mortality files and

We aggregate NCHS-based birth and death data for the production of national-level population estimates. When data by full geographic and characteristic detail are available, we use births directly as a base for the population under age 1. We apply death rates by characteristics and control to NCHS death data by aggregate characteristic groups (see the section on national estimates). For periods when full-detail data are not available, we use available or estimated data on vital events to calculate characteristic-specific rates which are then controlled to preliminary or provisional total data from NCHS.

Distributing the projected national-level births and deaths to the subnational level requires additional computations. To do this, we use a combination of short-term projections of county-level population characteristic detail and FSCPE data on the geographic distribution of total county vital events (where available). The projections are derived by calculating county-level age-specific fertility and mortality rates. We then apply these rates to the county population projections from the prior vintage. The resulting projected data by demographic detail are then reconciled with FSCPE data on the geographic distribution of total county vital events. These values are then summed to the state level and controlled to national projections of characteristics described above. The final county data are then controlled to the resulting state values. The national births and deaths serving as the controls for the subnational vital events have been adjusted to reflect the impact of the COVID-19 pandemic, resulting in an increase in deaths and decrease in births for numerous states and counties. Other than the national control, no adjustments were made to subnational birth or death estimates.

Net Domestic Migration

The third major component of the balancing equation is migration. Migration can be divided into net domestic migration (NDM) within the United States and net international migration (NIM) between the United States and elsewhere. The Population Estimates Program calculates domestic migration using several data sources and methods depending on the age group in question and the level of characteristic detail required.

For state and county total estimates, we calculate county-to-county net domestic migration based on four data sources:

1. Internal Revenue Service (IRS) tax return data for ages 0-64
2. Medicare enrollment data from Centers of Medicare and Medicaid Services (CMS) for ages 65+,
3. Social Security Administration's Numerical Identification File (NUMIDENT) for all ages
4. Change in the group quarters population (described in the "Group Quarters" section)

State and County Totals by Three Age Groups

We produce overall net rates of movement between counties for the total population estimates by three age groups: under 18, 18 to 64, and 65 and over. For the household population under age 18 and 18 to 64, we use person-level data on filers, spouses, and dependents from IRS tax return data. We match two years of IRS tax returns with age data from the NUMIDENT file to produce geographic data by age categories. The NUMIDENT is a database of all Social Security Numbers ever assigned, which is updated annually with new entries and any changes to a person's record.

2000-2010 Intercensal Population Estimates prepared by the United States Census Bureau. The life tables are for males and females in five groups: Hispanic, non-Hispanic White, non-Hispanic Black, non-Hispanic American Indian and Alaska Native, and non-Hispanic Asian and Pacific Islander.

Once tax returns are matched, we then compare the addresses between the two years of IRS data to identify the number of exemptions that moved from one county to another between tax filings. An IRS exemption is defined here as an individual who appears in the IRS tax return data, either as primary filer, spouse, or dependent.

Not all residents are represented in the tax exemption data, since not everyone files taxes. Therefore, the number of migrants in the IRS data is not equivalent to the number of migrants in the resident population. To overcome this coverage limitation, we calculate Net Domestic Migration (NDM) rates instead of using observed flows in the tax data. County specific net migration rates can be thought of as the ratio of net migrant exemptions to the number of exemptions present at the beginning of the migration period. Mathematically, the rate is first obtained by subtracting the number of out-migrants from the number of in-migrants for each county to produce the number net migrant exemptions. We then divide the net migrant exemptions by the sum of non- migrant and out- migrant exemptions for each county. We calculate these rates separately for each period by the two age groups (under 18 and 18 to 64), as follows:

$$NDM\ Rate_{0-17} = \frac{In\ migrants_{0-17} - Out\ migrants_{0-17}}{Non\ migrants_{0-17} + Out\ migrants_{0-17}}$$

$$NDM\ Rate_{18-64} = \frac{In\ migrants_{18-64} - Out\ migrants_{18-64}}{Non\ migrants_{18-64} + Out\ migrants_{18-64}}$$

Because the population aged 65 and over is more likely to enroll in Medicare than file taxes, we rely on Medicare enrollment data from CMS to account for movement of the older population. The process is similar to the under 18 and 18 to 64 age groups. Instead of tax exemptions, we match two years of Medicare enrollment data (address as of July 1) with age data from the NUMIDENT file. We then compare the addresses between the two years of Medicare data to identify the number of enrollees that moved from one county to another between the two years.

Similar to IRS filing, not everyone enrolls in Medicare. Therefore, the number of migrants in the Medicare data is not equivalent to the number of migrants in the resident population. For the same reason, we produce net rates based on Medicare enrollees for the 65 and older population. We calculate the net domestic migration (NDM) rate for the 65 and over population by subtracting the number of out-migrant enrollees from the in-migrant enrollees for each county to produce the number of Medicare-based net migrant enrollees. We then divide the number of Medicare-based net migrant enrollees by the sum of non-migrant enrollees and out-migrant enrollees for each county and period. The net rate is a ratio of the number of enrollees who moved in less those who moved out to the number of enrollees present at the beginning of the period, as given below:

$$NDM\ Rate_{65+} = \frac{In\ migrants_{65+} - Out\ migrants_{65+}}{Non\ migrants_{65+} + Out\ migrants_{65+}}$$

During the production of state and county total estimates, we apply these rates to the household population within the three age groups to produce a computed number of migrants for use in the balancing equation. We also treat change in GQ as an indirect measure of domestic migration. This methodology implicitly accounts for migration between GQ facilities as well as for household to GQ movement. To produce estimates of total migration for each of the three age groups, we combine age-specific domestic migration estimates from the application of these rates with the total amount of GQ population change in each age group. These total net domestic migration values are then controlled to sum to zero at the national level (as domestic migration must).

State and County Characteristics

The production of state and county characteristics estimates occurs after the production of state and county total estimates. The process for state and county total estimates only requires information on migration by age groups. However, to produce migration data by full characteristic detail, we need age, sex, race, and Hispanic origin.

To create net domestic migration estimates by full demographic detail, we use data from four sources:

1. Internal Revenue Service (IRS) tax return data for ages 0-64
2. Medicare enrollment data from Centers of Medicare and Medicaid Services (CMS) for ages 65+
3. Social Security Administration's (SSA) Numerical Identification File (NUMIDENT) for all ages
4. Demographic Characteristics File (DCF) for all ages

We use mailing address information from IRS tax return data for ages 0-64 to estimate migration. For ages 65 and older, we utilize address information from Medicare enrollment data to assign migration status. We use the NUMIDENT File to allocate age and sex to individuals in the migration universe.

The Population Estimates Program uses a Demographic Characteristics File (DCF) to allocate race and Hispanic origin to individuals in the migration universe with missing data. The DCF provides information on race and Hispanic origin. It is a dataset developed internally from a collection of person-level data derived from decennial census data, administrative records, and a set of imputation techniques when reported race and Hispanic origin are not available.

Because of known under coverage in the IRS and Medicare data (not everyone files taxes or claims benefits), we again calculate characteristic-specific out-rates and in-proportions and apply them to the population "at risk" of migrating. The population "at risk" is simply the population in each county in that particular age, sex, race, and Hispanic origin group.

We calculate domestic out migration rates by dividing the number of out-movers identified in the particular source data (IRS or Medicare, depending on age) by the total number of individuals at the beginning of the period. The total number of individuals at the beginning of the period is the sum of out movers and non-movers, as shown below:

$$Out\ Rate_{characteristic} = \frac{Out\ migrants}{Non\ migrants + Out\ migrants}$$

To distribute the pool of out movers by demographic characteristics to their destination counties, we use in-proportions. In-proportions are defined as the number of exemptions (for ages 0-64) or the number of enrollees (for ages 65+) moving into a county divided by the national total number of out mover exemptions/enrollees in a given demographic group. Though these can be very small proportions, this methodology creates in and out domestic migration rates that are consistent. It is important to note that no rounding is applied to these migration rates.

$$In\ Proportion_{characteristic} = \frac{In\ migrants}{\sum_{all\ counties} Out\ migrants}$$

In the production of state and county population estimates by characteristics, we apply the calculated out rates annually to each county's population "at risk" to produce estimated numbers of domestic out-migrants. Next, the national "pool" of out-migrants by demographic characteristics are then allocated to their destination counties with the in-proportions.

Net International Migration

The third major component of the balancing equation is migration. As noted, migration can be divided into NDM within the United States and NIM between the United States and abroad. We estimate international migration in several parts: immigration of the foreign born, emigration of the foreign born, net migration between the United States and Puerto Rico, net migration of natives to and from the United States, and net movement of the Armed Forces population to and from the United States. For each component, we first estimate the total migration flow for the nation.

For the sub-components of international movement other than the movement of the Armed Forces population, we use a proxy universe to distribute national total estimates by geography and demographic characteristic detail. A proxy universe is a geographic and characteristics distribution derived from a different population than the total estimate. We use a proxy universe because it allows us to utilize a larger sample (in the case of foreign- born immigration) and to produce characteristics for a population without direct observation (for native and foreign-born emigrants). We create total estimates for the United States either through direct or residual estimation, and then apply the distribution from the proxy universe in order to produce estimates of international migration for states and counties by demographic characteristics.

Again, excluding movement of the Armed Forces, state demographic characteristic distributions are based on three years of pooled American Community Survey (ACS) 1-year files, while county distributions are produced from ACS 5-year files. We control county-level data to state-level data to ensure the component data are consistent. For the net movement of the Armed Forces population, demographic characteristics and state distributions are developed based on a combination of data collected by the Defense Manpower Data Center (DMDC) and pooled 1-year ACS files.

We use the methodology described in the next sections to estimate the components of net international migration for April 2020 to June 2021. To account for the impacts of the COVID-19 pandemic, we adjust the estimate to reflect lower migration levels for the April 2020-June 2021 estimates period. We use the 2019 ACS and 2020-2021 monthly time series on international mobility from auxiliary data sources to estimate total net international migration for this period. Normally, we would use the 2020 ACS for estimating this period; however, the data are not available for use this vintage due to quality concerns (also resulting from COVID-19). Auxiliary data sources include data from the Department of Justice, the Institute of International Education, U.S. Citizenship and Immigration Services, and the U.S. State Department Bureau of Consular Affairs and Refugee Processing Center. Trends in auxiliary data between 2019 and 2021 are used to adjust 2019 ACS data to account for the COVID-19 pandemic, and applied to our foreign-born immigration, foreign-born emigration, and net native components. We use the proxy universe method to distribute characteristics for the nation, states, and counties. Because distributions from the ACS-based proxy universes lag by a few years, the geographic distribution and demographic composition of net international migration for 2020 will not reveal any localized COVID-19 effects.

Foreign-Born Immigration

We use the ACS residence one year ago (ROYA) question to estimate foreign-born immigration for the nation and states. We estimate foreign-born immigration separately for Mexico and All Other Countries since we expect these groups to exhibit different demographic and geographic patterns. We use the 1-year ACS to estimate national- and state-level totals. These totals represent international movement occurring between the previous year and survey year. For example, we would estimate movement between July 2021 and June 2022 from the 2022 ACS. Movement for the final year of the time series is equal to the previous year's estimate because the latest ACS lags behind the vintage year. For example, for Vintage 2022, we would use the 2021 ACS to estimate movement from July 2020 to June 2021 and hold the estimate constant for July 2021 to June 2022. We revise the estimate the following vintage when more current ACS data become available.

We use a proxy universe to distribute national and state characteristics as well as county totals and characteristics. The proxy universe for foreign-born immigration is the foreign-born population who entered the United States within five years of the survey. We adjust age to reflect age at year of entry. There are separate proxy universes for the Mexican born and for those born in another foreign country. We apply proxy universe distributions from pooled ACS files to the state-level totals to derive state characteristics. We aggregate state characteristics to derive national characteristics. Next, we apply proxy universe distributions from the 5-year ACS to state characteristics to derive county totals and characteristics.

Foreign-Born Emigration

We use a residual method to estimate emigration of the foreign-born population at the national level. The residual method uses information on mortality and recent immigration to account for cohort change in the foreign-born population within a specific period. Mortality estimates come from NCHS Hispanic life tables by age and sex. Immigration estimates come from the ACS year of entry question. We develop an annual time series from consecutive 1-year ACS files to measure foreign-born population change. We attribute to emigration (residual) any part of foreign-born population change not explained by mortality or immigration. Next, we divide the residual by person years to create annualized emigration rates. Finally, we apply the rates to the population "at risk" of emigrating by sex, year of entry, and place of birth cohorts to calculate annual foreign-born emigration totals for the nation. We calculate emigration rates for seven mutually exclusive groups: 1) Mexican-born males who entered the United States within the past 10 years, 2) Mexican-born females who entered within the past 10 years, 3) Mexican born who entered more than 10 years ago, 4) Canadian and European born who entered within the past 10 years, 5) Asian born who entered within the past 5 years, 6) All other foreign born who entered within the past 10 years, and 7) Asian born who entered more than 5 years ago and non-Mexican born who entered more than 10 years ago. We calculate separate rates under the assumption that each group exhibits different propensities to emigrate, as well as different demographic compositions and geographic distributions.

Using the first group as an example, we tabulate the Mexican-born male population who entered the United States within the past 10 years. Next, we survive this population forward to obtain the expected population for a later year. Subtracting the observed population from the expected population for the later year yields a residual, which is assumed to represent total emigration occurring over the period. Next, we convert this residual into an annualized emigration rate. We calculate six rates based on three 2-year residuals, two 3-year residuals, and one 4-year residual. In order to reduce the effects of survey variability, we average the six rates. We apply the averaged rate to the population at risk of emigrating (tabulated from the 1-year ACS) to obtain annual estimates of emigration.

We follow the same method for estimating emigration for the other six groups listed above. For groups (3) and (7), which represent non-recent arrivals, we average rates from multiple ACS files as an additional step to stabilize annual rates. These two groups have large "at risk" populations, and slight variability in emigration rates can cause improbably large fluctuations in the annual estimate of foreign-born emigration.

Migration between the United States and Puerto Rico

We use the ROYA question from the ACS and the Puerto Rico Community Survey (PRCS) to estimate annual migration flows between the United States and Puerto Rico. We classify ACS respondents who resided in Puerto Rico one year ago as in-migrants. We classify PRCS respondents who resided in the United States one year ago as out-migrants. We subtract out-migrants from in-migrants to calculate net migration. The proxy universe for net migration between the United States and Puerto Rico is the population born in Puerto Rico who entered the United States up to 10 years prior to the survey year. To account for the impact of COVID-19, we combine Bureau of Transportation Statistics (BTS) T-100 Airline Passenger Traffic Data (APT) with the 2019 and 2020 1-year ACS/PRCS for the April 2020-June

2021 estimates period. We continue to use ROYA from the ACS and PRCS to estimate migration for all other years, while flight data was also incorporated into our 2018-2020 estimates.[11]

Native-Born Migration

Estimates of net migration of the native-born population are produced using a method which utilizes data from approximately 80 countries. This work compares estimates of the United States-born or United States citizen population living overseas measured by population registers and censuses in other countries at two consecutive time periods. The residual is used to develop an average annual estimate of net native migration. The proxy universe for the net native migration component is the native-born civilian population whose residence one year ago was either in a different state or abroad (as this approximates the characteristics of people who migrate).

Movement of the Armed Forces Population to and from Overseas

We derive the estimate of the net overseas movement of the Armed Forces population from data collected by the Defense Manpower Data Center (DMDC). DMDC provides monthly tabulations of active duty military personnel stationed outside the United States by age, sex, race, Hispanic origin, and individual branch of service within the Department of Defense. We use a combination of DMDC and ACS data to estimate the population by race. We aggregate the DMDC data to four race groups: 1) White alone, 2) Black alone, 3) American Indian and Alaska Native alone, and 4) all other races. We then utilize ACS data to distribute the "all other races" group to the full-detail alone and in combination race groups needed for estimates production.

We assume that changes in the overseas military population, excluding deaths, indicate movement of personnel into and out of the United States. To derive estimates of net international movement of the Armed Forces at the county level, we primarily use DMDC data by age, sex, race, Hispanic origin. Five-digit zip code information from DMDC is matched to zip code information from the IRS to determine state and county location. When the DMDC zip code does not match, other information provided by DMDC is used to assign the state and county location. To improve the geographic distribution of the military population around certain domestic military installations, we use county grouping information derived from pooled 1-year ACS files.

National Population by Age, Sex, Race, and Hispanic Origin

The goal of the national population estimates process is to produce monthly resident population estimates by single year of age (0 to 100+), sex, Hispanic origin, and race (31 categories). We then divide these estimates into the following universes: household (HH), civilian (CIV), civilian noninstitutionalized (CNI), and resident plus armed forces overseas (RES+AFO). The core of the process is the demographic balancing equation. We take inputs on births, deaths, and net international migration by characteristics, and apply these components to the population at the beginning of the period.

The annual number of net international migrants is divided into monthly and quarterly values to use in the production of the estimates. The final year of available data (usually the year prior to the vintage year) is held constant to the end of the time series. Utilizing vital statistics (birth and death) information, however, is more complicated. Because we have limited final data by characteristics from NCHS, we use a combination of final and preliminary input data, populations at risk, rates, and controls. We use slightly different methods for three estimates periods based on data availability from NCHS: full-detail (up to two years prior to the vintage year), preliminary totals (the year prior to the vintage year), and no data (the vintage year to the end of the time series). Although the process across each of the three periods is essentially the same (apply rates to a population then control the result), the source of the rates and controls changes based on the level of detail available in the input data.

[11] For more detail on this component, refer to the section "Puerto Rico Resident Population by Age and Sex".

In the first period, from the base to two years prior to the vintage year, we have vital statistics by full characteristics. For births, we directly utilize the number of monthly events from NCHS data. For deaths, we multiply the starting population by life table death rates used in our Population Projections Program, and then control the result to NCHS deaths by sex and age (single-year-of-age under 70, and an aggregate of age 70 and over) for the following race and Hispanic origin groups: non-Hispanic White alone, non-Hispanic Black alone, non-Hispanic American Indian or Alaska Native alone, non-Hispanic Asian and Native Hawaiian or Other Pacific Islander alone, and Hispanic of any race.

In the second period, the year prior to the vintage year, we only have preliminary or provisional totals for births and deaths. Here we use the most recent year of final NCHS data (usually two years prior to the vintage year) to calculate characteristic-specific birth and death rates, apply those rates to the population, and control the result to the overall preliminary totals from NCHS.

In the third period, where NCHS data are not available (the vintage year), we use implied birth and death rates calculated from the most recent year for which preliminary data are available. We hold the preliminary totals constant, apply the new calculated rates to each period, and control to the annual total for every remaining year in the time series. From this point on, both the rates and the totals are held constant. Population size in each group provides characteristic-specific variation in the distribution of births and deaths.

There are three main steps in the production of monthly national population estimates (which include the vital statistics process above): estimate the quarterly national resident population; estimate the monthly population; and estimate the monthly population by the other four universes described earlier. The goal is to produce monthly resident population estimates by single year of age (0 to 100+), sex, Hispanic origin, and race (31 categories), then calculate the required population universes (e.g., household, civilian, etc.).

We create population estimates by quarter-years of age by applying final births, deaths, and international migration to the base, then aging the population forward one quarter-year of age. The process is repeated for every quarter in the time series. We round the final resident populations and components and assume any residual is part of international migration.

Once we have created final quarterly estimates of the population by characteristics, we estimate the population for the second and third month of each quarter. To do this, we assign the calculated monthly births and deaths for each quarter to specific months based on the monthly distribution of vital events in the most recent year of final NCHS data. Together with the international migration component, we use these vital statistics to estimate monthly values for population estimates by age, sex, race, and Hispanic origin.

The final step in the national estimates process is to calculate the additional population universes by demographic characteristics. To calculate the resident plus Armed Forces overseas population, we add the monthly overseas military population (from DMDC data) to the estimated resident population. The civilian population is the result of subtracting the monthly resident military population (also from DMDC) from the resident population. The civilian noninstitutionalized population is produced by subtracting the institutionalized group quarters population from the civilian population.[12] Finally, we estimate the household population by subtracting the total group quarters population from the resident population. In addition, we use linear interpolation to derive daily resident population estimates and monthly component settings (e.g., number of seconds per birth) for the Population Clock.[13]

[12] The institutionalized population is defined as people under formally authorized, supervised care or custody in institutions including correctional institutions, juvenile institutions, nursing homes, skilled nursing facilities, psychiatric hospitals, and facilities for the disabled.
[13] The Population Clock is published on the Census Bureau website and is located at http://www.census.gov/popclock.

State and County Total Resident Population

The goal of the state and county total population estimates process is to produce total population estimates and estimates of the state population aged 18 and over for all states, counties, and equivalents in the United States. We treat parishes in Louisiana, boroughs in Alaska, and several independent cities (in Maryland, Missouri, Nevada, and Virginia) as counties. The process focuses on the development of estimates for counties (and equivalents) only. State estimates exist only as a sum of the final estimates for counties.

Our process involves estimating the population separately for ages under 18, 18 to 64, and 65 and over. We estimate three age groups for this process for two reasons. First, we use different input data for domestic migration based on whether we are estimating a population under age 65 (IRS tax exemptions) or 65 and over (Medicare enrollment). Second, we produce estimates of the state population aged 18 and over to provide to the Federal Election Commission.

Producing state and county total population estimates is similar to the production of national estimates, as they are both based on the balancing equation. However, state and county estimates are produced for annual July 1 dates, and they incorporate domestic migration. Even though there are slight differences in the way we calculate the first three months (April to July) from the estimates base (using only one quarter of a year of migrants, for example), the process is very similar for all other points in the time series.

We first subtract the GQ population and "age" the population one year in order to produce an estimate of the household population at the start of each period. The aging process takes the proportion of the previous vintage county population age 17 and 64, applies that proportion to the current year, and moves that population into the next higher age group (e.g., the estimated number of 64-year-olds would "age" into the group aged 65 and over).

Net migration rates calculated from IRS and Medicare data are then applied to the aged household population at the start of the period to create estimates of net domestic migration. We then add net domestic migrants, add births (for the under 18 population), subtract deaths, and add international migrants to produce an uncontrolled estimate of the household population at the end of the period for each age group. The GQ population is then added to create uncontrolled resident population estimates for each age group.

The next step in the process ensures consistency with the national estimates. First, we control the calculated resident population numbers to equal the national numbers by the three age groups. Second, we add GQ change to the total household domestic net migration estimate for each age group and control that number to sum to zero at the national level by age group. We then round the final resident population by age group and allocate the remainder (usually very small) to the largest population value in the country. Finally, we aggregate the three age groups into total estimates for counties and sum these estimates to create final estimates for states.

State and County Resident Population by Age, Sex, Race, and Hispanic Origin

The goal of the state and county resident population estimates by demographic characteristics process is to create population estimates by age, sex, race, and Hispanic origin for all states, counties, and equivalents in the United States. This process essentially follows the cohort-component approach, adding births, subtracting deaths, adding the effects of net domestic and international migration, and aging the population forward. An additional factor in this process is the requirement of consistency between population estimates for the multiple levels of geography and characteristic detail (see the section on estimates consistency). County characteristics, for example, are produced by single year of age (under age 85 and an aggregate of age 85 and over), sex, Hispanic origin, and race (31 categories). Accounting for all the cross-classifications, there could potentially be 10,664 possible combinations per geographic area.

The calculation of state and county estimates by characteristics uses a two-way raking process to ensure that the final estimates sum correctly by both geography and characteristics. The method involves iteratively controlling estimated values to the larger geography's characteristics and the smaller geography's total estimates. In other words, we control state characteristics to national characteristics and state totals then control county characteristics to state characteristics and county totals. After repeated rakings, changes in the data become progressively smaller, eventually allowing us to round the result.

The raking process produces population estimates that are not necessarily integers. We then apply a controlled rounding process which allows us to convert the estimates to whole numbers without changing the total values. For state estimates, we control to both the state totals and the national characteristics. For county estimates, we control to the county totals and the final state characteristics. Because the state characteristic estimates have already been controlled and rounded, creating consistency between county characteristics and state characteristics automatically makes counties consistent with the national values as well.

Puerto Rico Resident Population by Age and Sex

The U.S. Census Bureau produces annual estimates of the resident population for the Commonwealth of Puerto Rico and its municipios. The estimates are produced by age and sex using a residual cohort-component approach as described previously for the United States, states, and counties.

The cohort-component population estimation method starts with the April 1, 2020 Blended Base population by age (0 to 99 and 100+) and sex, and then follows each birth cohort as it ages and experiences mortality and migration. This procedure is repeated for each year of the estimation period by age and sex. Final births and deaths data for each month from July 2019 to December 2019 and provisional births and deaths data for each month from January 2020 to June 2020 were obtained from the Puerto Rico Institute of Statistics and originate from the Puerto Rico Department of Health vital registration system. Natality data include month and year of birth, sex of the child, mother's age, and mother's country and municipio of residence; mortality data include month and year of death, sex, age, and decedent's country and municipio of residence.

United States/Puerto Rico migration was formerly estimated using the respondent's current residence and the ROYA question from the ACS and the PRCS. The migration estimates used in producing the annual Puerto Rico Commonwealth population estimates consisted of two components: in-migrants from the United States to Puerto Rico and out-migrants from Puerto Rico to the United States. Respondents to the PRCS living in Puerto Rico and indicating residence in the United States during the previous year were categorized as in-migrants. Respondents to the ACS living in the United States and indicating Puerto Rico residence during the previous year were labeled out-migrants.

In recent years, several major events have impacted migration patterns to and from Puerto Rico and necessitated methods to adjust survey-based estimates with flight data. These events include Hurricane Maria in September 2017, a 6.4-magnitude earthquake in January 2020, and the incessant effects of COVID-19 since March 2020. Starting in 2021, a flight-based methodology is used to calculate annual Puerto Rico Commonwealth total net migration using APT data from BTS. Summarizing monthly APT data to the estimates year (July-June) and including all flight segments for Puerto Rico produces reasonable results. Moving to a flight-based method is advantageous as it improves the accuracy and recency of net migration estimates for Puerto Rico and reduces the number of future adjustments needed to account for major events impacting migration.

Annual net migration totals are calculated using in-bound and out-bound passenger flow data for the estimate year (EY) using APT data: in-bound total passengers to Puerto Rico are subtracted from out-bound total passengers from Puerto Rico to obtain the net migration estimate. Domestic and international flight data are used in the estimates. International flight data account for movement from abroad. Domestic flight data are received every 3 months while international flight data are received every 6 months. However, the Census Bureau has an internal agreement with BTS to access international data with only a 3-month lag, enabling us to include international flights in our estimates.

The ACS/PRCS is still used to calculate demographic characteristics. Sex is tabulated using 1-year PRCS (ROYA) data for in-migration flows, and then distributed to single years of age (1-114, 115+) using within-sex proportions from PRCS (ROYA) 5-year estimates. This process is repeated to determine out-migration flows using the ACS (ROYA) 1-year estimates for sex and the ACS (ROYA) 5-year estimates for age proportions within sex. A spline regression is used to smooth single years of age within sex. Lastly, raking and rounding procedures are implemented to generate the final net migration estimate by sex and single year of age (0-99, 100+).

Seasonal pattern variation related to tourism, particularly during the summer and winter months, may lead to trends that require annual fluctuation adjustments of flight data. Net migration this EY showed atypical positive net passenger flows to Puerto Rico from the United States for the months of March and June 2021. Data anomalies coincide with seasonal pattern variation potentially stemming from the lifting of travel restrictions in March and an early start to the tourism season in Puerto Rico. An adjustment was necessary in order to address this early seasonal pattern variation.

To make our adjustment, a simple average was obtained between the EY 2019 and EY 2021 flight data for March to June 2021. We selected EY 2019 for averaging as this was the last year prior to the earthquake and COVID-19 pandemic, and it was not impacted by Hurricane Maria. Data for 2021 also showed that seasonal trends returned to 2019 patterns but at elevated levels due to travel restrictions lifted in March 2021.

Net migration for Puerto Rico municipios was estimated using a residual method. The expected population for each municipio on April 1, 2010 was projected from the Census 2000 count by accounting for change since that census due to births and deaths. Differences between the expected population and the population enumerated in the 2010 Census are assumed to represent net migration over the decade. The residual, which represents events over a period of 10 years, was converted into an annual average migration rate by age and sex for each municipio. These rates are then controlled to the national rate. This method will be updated in upcoming vintages as newer data become available.