METHODOLOGY FOR THE UNITED STATES POPULATION ESTIMATES: VINTAGE 2022
Nation, States, Counties, and Puerto Rico – April 1, 2020 to July 1, 2022

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 (ACS), for community development, to aid business planning, and as denominators for statistical rates, among many other uses. Overall, the estimates time series from 2010 to 2020 was very accurate, even accounting for ten years of population change. The mean absolute percent error (MAPE), which is the average absolute difference between the final total resident population estimates and 2020 Census counts, was only about 2.9 percent across all counties.[2]

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

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 available at https://www.census.gov/programs-surveys/popest/technical-documentation/methodology.html.
[2] For more information on estimates evaluation, see https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates.2020.html#list-tab-J3V9IOEBY6N8ZHN9W1.
[3] Domestic migration sums to 0 at the national level and therefore has no effect on the estimates.

Estimates Consistency, Controlling, and the Residual

The estimates are produced 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. The difference between the result of the balancing equation and the final estimate is referred to as 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. 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 single 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, these data are rounded 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 unique challenges.

For example, the disclosure avoidance system applied to the 2020 Census counts impacted 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).

Additionally, the COVID-19 pandemic introduced significant delays to both the enumeration of the 2020 Census and post processing of the data, affecting the availability of official decennial data by the full age, sex, race, and Hispanic origin detail required for estimates processing relative to the annual estimates production schedule.

In response, 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 was created for the Vintage 2021 series of estimates, reflecting the most detail from alternate sources we could confidently incorporate into the estimates base with the time that was available to finalize the method. This approach was retained for the current Vintage 2022 series featuring a limited number of modifications. The Vintage 2022 blended base consists of the following:

- **2020 Census Data:** 2020 Census data from the internal Census Edited File (CEF) tabulated into 2022 geographies at the subcounty level, infused with differentially private noise, and then aggregated to create county, state, and national total resident, household, and group quarters (by facility type)[5] population counts
- **2020 Demographic Analysis (DA)[6] Estimates**: National population estimates by age and sex
- **Vintage 2020 Population Estimates for April 1, 2020 (V2020)**: Nation, state, and county population estimates by age, sex, race, and Hispanic origin

The blended base process for Vintage 2022 began with tabulating population counts on the internal 2020 CEF by 2022 geographic boundaries at the subcounty level. These counts were infused with a small amount of differentially private noise[7] before they were aggregated national, state, and county levels for use as population controls. The demographic characteristics (age, sex, race, and Hispanic origin) from the V2020 estimates for April 1 were applied to the protected 2020 Census totals by geography. This process created an initial dataset with aggregated 2020 Census totals featuring V2020 characteristics, hereafter referred to as "census-adjusted V2020."

From there, the blended base process incorporated the 2020 DA estimates using a top-down methodology which is very similar to the process for developing the postcensal population estimates each year, described earlier in this document: first, we created blended national-level data by raking 2020 DA to national controls created from the protected 2020 Census data. Then, the census adjusted V2020 estimates were raked by single year of age and sex to the nationally controlled DA data. This integrated the protected 2020 Census counts, 2020 DA estimates, and April 1, 2020 estimates from V2020 into the finalized national blended base population. At the national level, then, it is accurate to say that resident, household, and group quarters (GQ) population totals are derived from the 2020

[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 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."

[6] 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

[7] For more information regarding confidentiality protections applied using the 2020 Census Disclosure Avoidance System, see https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance.html.

Census, age and sex detail is drawn from 2020 DA, and race and Hispanic origin detail comes from V2020.

Next, we raked the census-adjusted V2020 state-level estimates to the national level blended base by full detail and to the protected 2020 Census state totals. This allowed us to retain the benefits of the national blended base while maintaining consistency with the aggregated protected 2020 Census data. We developed the county-level blended base data using a parallel method, raking the census-adjusted V2020 county-level data (in Vintage 2022 geographic boundaries) to the state blended base and the protected 2020 Census population totals for counties. Finally, we rounded, aggregated the county-level estimates to ensure geographic consistency, and modeled additional detail required for our estimates processing (e.g., quarter years of age) using the V2020 data.

The blended base for Puerto Rico is developed using a simplified version of the process described above. Since these estimates are produced for the Commonwealth and municipios by age and sex, and not for lower levels of geography, capturing geographic changes was not a factor. Additionally, there is no DA control available for Puerto Rico. The Vintage 2022 blended base for Puerto Rico consists of the following:

- **2020 Census Data**: Commonwealth and municipio total resident, household, and GQ population counts from the PL 94-171 Redistricting File
- **V2020 Population Estimates for April 1, 2020**: Commonwealth and municipio population estimates by age, sex, and universe (including GQ facility type)

The Puerto Rico Commonwealth blended base was developed by raking the V2020 April 1, 2020 population by age and sex directly to the 2020 Census total counts. Municipio data then followed 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 GQ population every year by single year of age, sex, race, Hispanic origin, and facility type. The GQ method begins with the April 1, 2020 blended base GQ population. 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 April 1, 2020 blended base 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,

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 used in Vintage 2022 were from calendar year 2020. Additionally, for Vintage 2022 we utilized the available NCHS provisional data to account for recent trends and COVID-19 impacts on natality and mortality, which varied in recency by component. For births, we had NCHS monthly provisional totals by state for 2021 through March of 2022. For deaths, we had NCHS monthly deaths with some characteristic detail for 2021 through June 2022 at the national level and monthly deaths by state for the same time period. The NCHS provisional data are used in conjunction with the data received from the FSCPE to approximate final NCHS data with characteristic assignment coming from the last year of final NCHS data when information on demographic characteristics is otherwise unavailable.

We also modify the NCHS births and deaths data to comply with our process. The first change to births is to assign race of child. Birth certificates include only data on the race and Hispanic origin of the parents, not the child, so 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.

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

While we make no 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]

We aggregate NCHS-based birth and death data for the production of national-level population estimates and 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).

Distributing the national-level births and deaths to the subnational level requires additional computations. We distribute the monthly state provisional totals for 2021 and 2022 using characteristic and geographic distributions from the last year of final data (2020). These data are reconciled with the FSCPE data on the geographic distribution of total county vital events. These values are then summed to the state level and controlled to the national births and deaths by characteristic described above. The final county data are controlled to resulting state values. The national births and deaths serving as controls for the subnational vital events utilize recent provisional data to reflect the impact of the COVID-19 pandemic and other recent changes in natality and mortality trends. No other adjustments were made to the subnational birth or death estimates.

[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 the United States Census Bureau's 2017 National Population Projections. 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.

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-specific net domestic migration based on four data sources:

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

State and County Totals by Three Age Groups

We produce overall net rates of movement into and out of each county 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.

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-over population. We calculate the 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 to 64
2. Medicare enrollment data from Centers for Medicare and Medicaid Services (CMS) for ages 65 and older
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 to 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. 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 to 64) or the number of enrollees (for ages 65 and older) 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, it is important to note that no rounding is applied before calculating migrants.

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

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 allocated to their destination counties with the in-proportions. As a result, the in and out domestic migrants across all counties naturally aggregate to zero at the national level.

Net International Migration

As noted, the third major component of the balancing equation is migration: NDM within the United States and NIM between the United States and abroad. We estimate international migration via several components: non-U.S.-born immigration, non-U.S.-born emigration, net migration between the United States and Puerto Rico, net migration of the U.S.-born population 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 migration totals for the nation. Then, we use a proxy universe to distribute totals (except net movement of the Armed Forces) into single year of age, sex, race, and Hispanic origin for the nation, states, and counties. For the net movement of the Armed Forces population, demographic characteristics and state and county distributions are developed based on a combination of data collected by the Defense Manpower Data Center (DMDC) and pooled 1-year ACS files.

A proxy universe is a subset of the total ACS population which we assume will represent the actual geographic and demographic composition of international migrants. We use proxy universes because the sample size is considerably larger than the input data we use for estimating national migration totals. In addition, a proxy universe provides characteristics of migrants that we cannot directly measure, such as emigrants. National migration totals are distributed to national and state characteristics using a proxy universe pooled from three ACS 1-year files. National and state characteristics are then distributed to counties using a proxy universe based on ACS 5-year data. Using Vintage 2018 as a hypothetical example, this would mean we would distribute the 2017 national migration totals from pooled 2015-2017 ACS 1-year files and the 2017 ACS 5-year file. Since the ACS data

release lags one year behind the vintage year, the 2017 distributions would be held constant for 2018. We would then update the 2018 distributions in Vintage 2019 with the 2018 ACS files and hold the distributions constant for 2019.

COVID-19 Adjustments

After Vintage 2019, we updated the methodology to account for the impact of the global COVID-19 pandemic on international migration. The national non-U.S.-born immigration, non-U.S.-born emigration, and U.S.-born net migration totals are adjusted according to 2019-2022 time series on visa issuances, new student enrollments, refugee admissions, and humanitarian migrant cases from the Bureau of Consular Affairs, Institute of International Education, Refugee Processing Center, U.S. Citizenship and Immigration Services, and Department of Justice. COVID-19 adjustments are for national totals only. We make no adjustments to states, counties, and characteristics.

The national-level adjustments are as follows:

> National migration totals for 2020 (July 1, 2019 - June 30, 2020) are set to 76% of 2019 levels. About 2.6% of the annual total is allocated into the quarter year (April 1, 2020 - July 30, 2020) estimate.
>
> National migration totals for 2021 (July 1, 2020 - June 30, 2021) are set to 40% of 2019 levels.
>
> For 2022 (July 1, 2021 - June 30, 2022), only non-U.S.-born immigration is adjusted. National migration total is set to 103% of 2019 levels. The other components are computed using our regular methodology.
>
> Net Puerto Rico migration totals for 2020 and 2021 are retained from Vintage 2021. Adjustments are based on T-100 airline passenger traffic data from the Bureau of Transportation Statistics combined with previous 1-year ACS and Puerto Rico Community Survey (PRCS) migration (residence 1 year ago) data.

The next sections describe the regular NIM methodology by component.

Immigration (Non-U.S.-Born)

The immigration component accounts for international moves into the United States in which a change of usual residence has occurred, regardless of citizenship or migrant status. This includes the foreign born (naturalized citizens and non-citizens) and persons born abroad to U.S. citizen parents. We use ACS 1-year migration (residence 1 year ago) data to estimate national and state immigration totals. For example, for Vintage 2018 we estimated immigration totals for 2017 (July 1, 2016 - June 30, 2017) from the 2017 ACS 1-year file. Since the ACS data release lags one year behind the vintage year, we held the 2017 immigration totals constant for 2018. We then updated the 2018 totals in Vintage 2019 with the 2018 ACS 1-year file and held these values constant for 2019.

Immigration totals are estimated for the following groups:

1. Mexico
2. All Other Countries

The proxy universe for immigration (Mexico) is the ACS population born in Mexico whose year of entry was five years ago or less. The proxy universe for immigration (All Other Countries) is the ACS population born in a foreign country (except Mexico) whose year of entry was five years ago or less. We adjust the age distribution to reflect age at arrival in the United States.

Emigration (Non-U.S.-Born)

The emigration component accounts for the resident foreign born (naturalized citizens and non-citizens) and persons born abroad to U.S. citizen parents that moved out of the United States for whom a change of usual residence has occurred. We use a residual method with ACS 1-year files to estimate emigration at the national level. The residual method calculates implied emigration (residual) by subtracting deaths and recent immigration from population change measured between two ACS 1-year files. We apply survival ratios to the non-U.S.-born population from the first ACS file to estimate period deaths. Survival ratios are constructed from annual NCHS Hispanic life tables by single year of age (0-99, 100+) and sex. We use ACS migration (year of entry) responses from the second ACS file to estimate immigration that occurred since the first ACS file. The residual is the remaining change in the ACS population after accounting for deaths and recent immigration. Residuals are calculated separately for the following mutually exclusive groups:

1) Males born in Mexico who entered the United States within the past 10 years
2) Females born in Mexico who entered within the past 10 years
3) All persons born in Mexico who entered more than 10 years ago
4) All persons born in Canada or Europe who entered within the past 10 years
5) All persons born in Asia who entered within the past five years
6) All persons born in another country who entered within the past 10 years
7) All persons born in Asia who entered more than five years ago, or born in another country (excl. Mexico) who entered more than 10 years ago

We assume each group will exhibit different propensities to emigrate and different characteristics. These groups are also the definitions for the proxy universes constructed from the ACS files.

We calculate three versions of residuals using the 2010-2014, 2010-2015, and 2010-2016 ACS periods. We divide the residuals by person-years to get annualized rates. We average the rates from the three versions to produce the final emigration rate by group. We apply the rate to the previous ACS 1-year population by group to estimate national emigration totals for the current year. We use proxy universes to distribute national totals to states, counties, and characteristics.

Migration between the United States and Puerto Rico

We use ACS and PRCS 1-year files on migration (residence 1 year ago) to estimate annual migration flows between the United States and Puerto Rico. We subtract PRCS respondents who resided in the United States one year ago from ACS respondents who resided in Puerto Rico one year ago to calculate net migration into the United States. The proxy universe is the ACS population born in Puerto Rico who entered the United States within the past 10 years.

This component only accounts for migration between the United States and Puerto Rico. For all migration to and from Puerto Rico, please see the section on Puerto Rico Resident Population by Age and Sex.

U.S.-Born Net Emigration

We use a residual method on census, survey, and population register data from approximately 100 countries to estimate net migration of the population born in the United States. The residual method measures change in the population of Americans (born in United States or citizens) in foreign countries measured at two points in time. We assume the change in population after accounting for deaths represents net migration (residual) between the United States and foreign country. The residual is divided by the number of years between the two data points to produce an annualized country total. Country totals are then aggregated to a global total, which we hold constant every year.

The estimate is periodically updated as current foreign data become available. The proxy universe is the civilian population born in the United States whose residence one year ago was either in a different state or abroad.

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 DMDC. The 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 indicate movement of personnel into and out of the United States. The national-level estimates of net international movement of the Armed Forces are distributed to counties using the DMDC active-duty military population residing in the United States by age, sex, race, and Hispanic origin as a proxy universe. 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. For births, we directly utilize the number of monthly events from the births described above. For deaths, we multiply the starting population by life table death rates used in our Population Projections Program, and then control the result to the deaths described above 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.

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 for the other four universes described earlier.

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 deaths for each quarter to specific months based on the monthly distribution of deaths in the 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 GQ population from the civilian population.[11] Finally, we estimate the household population by subtracting the total GQ population from the resident population. In addition, we use linear interpolation to derive daily resident population estimates and monthly component settings (e.g., average occurrence of births, deaths, and migration in seconds) for the Population Clock.[12]

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, planning regions in Connecticut, 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. To produce an uncontrolled estimate of the household population for each age group at the end of the period, we start with the aged household population and add births (for the under 18 population), subtract deaths, add net domestic migrants, and add

[11] 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.
[12] The Population Clock is published on the Census Bureau website and is located at http://www.census.gov/popclock.

international migrants. 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 years of age (0-84 and 85+), sex, Hispanic origin, and race (31 categories), resulting in 10,664 characteristics 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 multiple raking iterations, 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. We produce the estimates by age and sex using a cohort-component method as described previously in the overview of the document.

The cohort-component population estimation method starts with the April 1, 2020 blended base population by age (0 -100+) and sex, and then follows each birth cohort as it ages and experiences mortality and migration. We repeat this procedure for each year of the estimation period by age and sex. We obtain the most recent final births and deaths microdata by month for calendar years 2019 through 2021 and provisional births and deaths data for each month from January 2022 to June 2022 from the Puerto Rico Institute of Statistics. These data 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, mother's country and municipio of residence; mortality data include month and year of death, sex, age, decedent's country and municipio of residence.

In recent years, several major events impacted migration patterns to and from Puerto Rico and necessitated methods to adjust survey-based migration estimates with flight data. These events include Hurricane Maria in September 2017, a 6.4-magnitude earthquake in January 2020, and the effects of COVID-19 since March 2020. We started using a flight-based methodology to calculate annual Puerto Rico Commonwealth total net migration with Airline Passenger Traffic (APT) data from the Bureau of Transportation Statistics (BTS) in Vintage 2021. We summarize monthly APT data for each estimate year (EY), July 1 to June 30, and include all flight segments for Puerto Rico. Moving to a flight-based method improved the accuracy and recency of net migration estimates for Puerto Rico and reduced the number of future adjustments needed to account for major events impacting migration.

We use both domestic and international flight data to create net migration estimates, with international flight data accounting for movement from abroad. We calculate annual net migration totals by using in-bound and out-bound passenger flow data for the EY. We subtract total in-bound passengers to Puerto Rico from total out-bound passengers from Puerto Rico to obtain the net migration estimate. Each vintage, we revise the previous year's net international migration number using the final BTS data which became available the following year.

Seasonal pattern variation related to tourism, particularly during the summer and winter months, may lead to atypical trends that require annual adjustments of flight data. Data anomalies coincide with seasonal pattern variation potentially stemming from the lifting of COVID-19 travel restrictions and varying pandemic phases. A simple adjustment was necessary to address early seasonal pattern variation in Vintage 2021. A similar adjustment was implemented in Vintage 2022.

To make our EY 2022 adjustment, a simple average was obtained between EY 2019 and EY 2022 flight data for July-August 2021. We selected EY 2019 for averaging as this was the last year prior to the 2020 earthquake and COVID-19 pandemic, and it was not impacted by Hurricane Maria. Data for 2021 also showed trends returned to 2019 patterns, but at elevated levels due to travel restrictions lifted in March 2021, continuing through the summer.

The ACS/PRCS is used as a proxy universe to obtain age and sex characteristics for migrants. For in-migration, sex is tabulated using one-year PRCS residence one year ago (ROYA) data. This is then distributed to single year of age (1-115+) using within-sex proportions from PRCS (ROYA) five-year estimates. This process is repeated to determine out-migration while using the ACS (ROYA) one-year estimates for sex and the ACS (ROYA) five-year estimates for proportions within sex. A spline regression is applied to smooth single years of age within sex. Single years of age are then collapsed (1-100+), and for those younger than one year old, age is imputed to half of age one. Next, to acquire the "pre-controlled net migration estimate", out- migration and in -migration from the previous step is subtracted. Afterward, the APT annual net migration total is divided by the "pre-controlled net migration estimate" to get the rake factor. For the "controlled in- and out-migration estimates" by sex and single year of age (0-100+) the rake factor is applied to the pre-controlled in- and out-migration estimates. Lastly, for the final net migration estimate by sex and single year of age, we apply greatest mantissa rounding to the "controlled" in- and out-migration estimate, and then subtract in- and out-migration.

We use migration rates as initial inputs for Puerto Rico municipio estimates. We calculate these rates using a residual method, where differences between the expected population and the enumerated population represent net migration over time. We rake municipio estimates to Commonwealth estimates and characteristics. The residual estimate (preliminary - final PRM estimates) and GQ change is assumed to be part of net migration. Starting in Vintage 2022, these municipio net migration rates are now calculated for the 2010 to 2020 period using 2010 Census counts and Vintage 2021 blended base estimates. This is not a method change, but rather an incorporation of more recent data. The new net migration rates have a small-to-medium impact on most municipios.