## PAA 2015 Paper/Poster Proposal: Short Abstract

***Title:*** *Processing the Census Bureau's Population Estimates*
***Authors:*** *Ben Bolender and Larry Sink, Population Estimates Branch*

The Population Estimates and Projections area at the U.S. Census Bureau develops population estimates for the nation, states, and counties by age, sex, 31 race groups, and Hispanic origin; population estimates for Puerto Rico and its municipios by age and sex; and total population estimates for cities, towns, and school districts. These estimates are the foundation for funding allocations, survey controls, and the development of federal, state, and local government agency statistics. To develop and release these estimates by mandated deadlines, we have created a sophisticated process for data acquisition, simulation, processing, review, and dissemination. This project describes our annual schedule, input data sources, matrix-oriented workstructure, teams and internal processes, systems for sharing data internally, and methods we use to release our data to interested data users and the public. It provides a clear overview of the organization behind our estimates production process.

## PAA 2015 Paper/Poster Proposal: Long Abstract

***Title:*** *Processing the Census Bureau's Population Estimates*
***Authors:*** *Ben Bolender and Larry Sink, Population Estimates Branch*

**Introduction**

The Population Estimates and Projections area at the U.S. Census Bureau develops population estimates for the nation, states, and counties by age, sex, 31 race groups, and Hispanic origin; population estimates for Puerto Rico and its municipios by age and sex; and total population estimates for cities, towns, and school districts. These estimates are the foundation for funding allocations, survey controls, and the development of federal, state, and local government agency statistics.

Production of the estimates requires a variety of input data. The demographic balancing equation is relatively simple. Population change is a result of births, deaths, or migration (the movement of people from one area to another).

*Figure 1. The Balancing Equation*


In the real world, however, there are no perfect or easily-accessible sources for these inputs with sufficient detail to allow for the production of estimates in the specific categories we use. Instead, we rely on a variety of administrative and survey-based input data. The incoming data do not always fit the demographic categories, geographies, or time periods we require. Because of this, we have developed a variety of processes and structures to utilize the data we receive to produce the population estimates our data users and the public need.

This project presents a high-level overview of the structure of population estimates production at the U.S. Census Bureau. We begin by describing the overall process flow in terms of our production schedule and output data requirements. We then list the various input data that we use for each of the major components in the balancing equation (and other distributive processes). This includes a list of the federal and state agencies that provide data to the Census Bureau for estimates production along with a brief review of the level of detail associated with each request.

We move then to a discussion of the simulation and production process. This includes our change control process, requirements and internal data requests, research and simulation, production, internal review, audit, and data delivery. Next, we describe the two major avenues we use to release our data to the public and interested users. On the one hand, our dissemination process creates public use tables and files which are posted on the Census Bureau's Web site, and creates materials for the media. On the other, we provide special tabulations to internal Census Bureau surveys, other government agencies, and interested data users to allow more precise tabulations than would be possible with our publicly available tables.

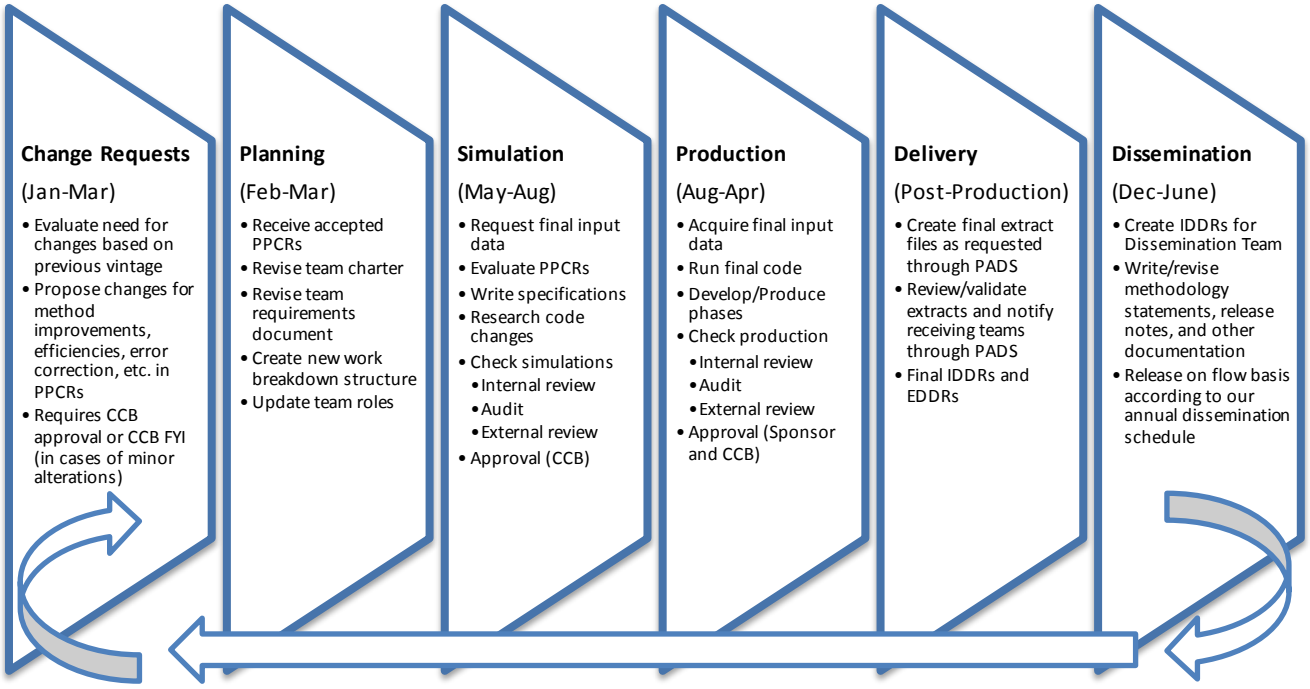
Finally, related to our dissemination activities, we will briefly discuss the work that goes into creating our media and press release materials. In addition to the data, we are continually reaching out to the public and other agencies to provide information, support, and details regarding our data. These include press releases, media tours, conference presentations, blog posts, graphics, and telephone data consultation. Overall, the high degree of organization in our processes allows us to produce a complete data package to inform people about the nation's people and places.

**Production Timeline and Requirements**

In general, we produce our estimates time series in "vintages" which start at the previous decennial census and estimate forward using measures of population change over the entire period. For example, our Vintage 2013 estimates recreated the time series from April 1, 2010 through July 1, 2013. Our annual production schedule is driven by a few major deadlines. We publicly release total population estimates for the nation, states, and Puerto Rico every year in late December. The Federal Election Commission requires our estimates of the population in each state aged 18 or over by early January. In March, we release data on the total population of counties, county equivalents, and Puerto Rico municipios, along with components of change and aggregated estimates for Core-Based Statistical Areas (metropolitan and micropolitan areas). Between January and April, we deliver survey controls to the Current Population Survey and the American Community Survey, among other data users. May sees the release of our subcounty total estimates for cities and towns, and housing unit estimates for counties. Finally, in June we publicly release data on the nation, states, and counties by age, sex, race, and Hispanic origin along with estimates for Puerto Rico and its municipios by age and sex.

In order to accomplish the release of these products, we have created a standard schedule for use throughout the vintage. The following flow chart illustrates a typical year in our production time line.

*Figure 2. Population Estimates Overview Flowchart*




Note: PADS is our internal system for moving data between teams.
Abbreviations: PPCR=Proposal and Production Change Request; CCB=Change Control Board; IDDRs=Internal Detailed Data Requests EDDRs=External Detailed Data Requests

Every year, our teams propose changes, plan for production, simulate change, request new input data, produce final population estimates, review the results, and deliver estimates data products both internally and to data users outside the estimates area. Since our time line has an overlapping start and finish, we are consistently multitasking data acquisition, processing, and dissemination tasks.

**Input Data**

The input data we use come from a variety of administrative and survey-based data sources. All of our estimates start with the population base, or the population at the time of the last decennial census. For our Vintage 2013 estimates, the base comes from Census 2010. The original census values are adjusted using Count Question Resolution program revisions, any geographic changes that were incorporated since the census date, and the results of other Census operations. Further, we modify race categories to remove the "Some other race" group for estimates processing. We do this to produce estimates in the five Office of Management and Budget race categories and their combinations (White; Black or African American; American Indian or Alaska Native; Asian; and Native Hawaiian or Other Pacific Islander).

Data on births are received from the National Center for Health Statistics (NCHS) as individual birth records with information on the child, mother (and father when available), and the place of residence. NCHS provides similar individual death records including age, sex, race, Hispanic origin, and residence. One processing complication is that some states still record birth and death data in the 1977 Office of Management and Budget race categories (four race groups), and we need to convert these data into the 31 race groups we produce. Another is that final data by full detail is only available for the period two years prior to the vintage year. This means we need to create a short term projection in order to produce estimates. We supplement these data (especially in the projection period) with information on county distributions provided by the Federal-State Cooperative for Population Estimates (FSCPE) members.

Data on international migration are derived from a variety of sources, including the American Community Survey (ACS), the Puerto Rico Community Survey (PRCS), research on the emigration of the native population, and data from the Defense Manpower Data Center (DMDC) on the movement of military personnel between the United States and other areas. While we have internal access to the ACS and PRCS, data from the DMDC must be acquired externally. The DMDC data also provide information on the military population that we use to create estimates of the civilian and civilian noninstitutionalized populations, which are often used as survey controls.

Domestic migration rates are created from a combination of Internal Revenue Service (IRS) tax exemptions, Medicare enrollment, the Social Security Numeric Identification file (NUMIDENT), and our internal Demographic Characteristics Database (DCDB). Through a special agreement (and under Title 26 which governs our use of tax data), we are provided information on tax exemptions which we match across years to determine migration rates for various counties. The NUMIDENT provides information on migrant and non-migrant age and sex, while we derive information on race and Hispanic origin from the DCDB, chiefly based on decennial census data.

We also receive estimates of change in the group quarters population at the facility level from our state partners in the FSCPE through the annual Group Quarters Report (GQR). We use these data in our estimates to create estimates of the household population, the civilian noninstitutionalized population, and as part of our published measures of domestic migration. We get information on housing unit change from the Building Permit Survey, Survey of Construction, Manufactured Homes Survey, and the American Housing Survey, all conducted by the Census Bureau.

**Team Structure and Processes**

The production of the estimates requires a highly skilled and committed staff along with a high degree of organization and coordination. We have met these challenges by integrating staff into a variety of purpose-focused teams with clearly defined processes and scope. Some teams are tasked with the acquisition of external input data. Others clean and process these data to prepare them for production. Production teams maintain the methodology that combines the various inputs into final estimates and ensures that these estimates are internally consistent across time periods, geographies, and demographic characteristics. Review teams check simulation and production data for demographic reasonableness, differences relative to the last vintage, and differences due to methodological change. Finally, dissemination and external data request teams work with our management, our Public Information Office, the media, and external data customers to ensure our data is available to the widest possible audience.

Even with these varied tasks, all of our teams function with a similar structure following project management principles. Each team has a charter, team requirements document, and work breakdown structure which clearly lay out the team's requirements, tasks, acceptance criteria, and relationship with both other teams and the production schedule as a whole. Within each team, we also have similar roles, which allow staff to be flexible as preferences allow or schedules dictate. This works well with our matrix-oriented structure, where staff across the area work on a variety of teams, regardless of their actual Branch location. Every team has a team leader, responsible for organizing the team activities and completing work tasks on time. The team leader is also generally responsible for representing the team at meetings or as an expert for consultation. The developer writes and maintains data production code while maintaining change control. Reviewers are tasked with examining the data for reasonableness and correctness while auditors check the processing code for errors or extraneous changes. Other roles on the team also provide for passing data between teams, checking data extracts, creating special tabulation files for release, and checking that all documentation is complete and updated.

**Simulation, Production, and Review**

During the simulation season, teams work to develop improvements to estimates methodology, write new code to incorporate them, review their potential effects on the estimates, and have them approved by the Change Control Board. This may be as simple as a single input team running a single simulation, or as complicated as our "monster" simulation at the end of the summer, which completely recreates the previous vintage of estimates including all of the data transfers, documentation, and review. The reason we do this is to test the combined impact of all of that year's proposed changes, paying particular attention to any interaction effects. Once this "monster" simulation is complete, we then move to the production phase where we acquire, review, and incorporate the new input data requested over the summer. Using the approved methodology from simulations, we apply our production process to these new inputs to create the final series of population estimates.

To put the scale of this task in perspective, the final datasets include: monthly national estimates by 101 ages, sex, 31 races, and Hispanic origin; annual state and county estimates by three age groups (0-17, 18-64, 65 and over); annual state and county characteristics by 86 ages, sex, 31 races, and Hispanic origin; annual Puerto Rico commonwealth and municipio estimates by 86 ages and sex; annual total population estimates for cities, towns, and all the pieces of subcounty geography needed to create them (approximately 80,000 records); annual housing unit data for counties; and all the input data sets, at similar levels of detail, needed to produce them. For reference, the final county characteristics dataset has the potential for 10,664 records (age/sex/race/Hispanic origin combinations) per county and a total of approximately 33.5 million records per year of the time series.

These estimates are reviewed heavily for accuracy and consistency before they are released. First, the production team reviews the data to check for processing errors and high level changes. Then, the external review team ("external" meaning outside the production team) examines the data in terms of reasonableness and vintage-to-vintage change. Incorporating "big" data on administrative records and surveys, along with the complexity of our process can lead to occasional anomalies in the data. The goal of the external review process is to identify any anomalies that exist, determine their cause, and make recommendations on whether adjustments or corrections to the data are needed.

**Dissemination and External Data Requests**

Once the data are reviewed and the results approved by management, we move on to the dissemination phase of the process. This takes two primary forms. The first is the public release of our data on Census Bureau Web sites. The second is through special tabulations and data requests. Both are key components of our mission to produce and disseminate information on the nation's people and places.

The dissemination process begins by developing table packages, release notes, methodology statements, and related material for public release. We then work with management and other areas of the Census Bureau to publish our data on sites like American FactFinder and our Population Estimates site. We also work closely with our Public Information Office to develop reference materials, reach out to the media, conduct radio and television interviews, and answer reporters'

questions about the data we released. While these activities peak around our major releases, we are continually consulted throughout the year as experts on population estimates, population change, and demographic methodology.

The external data request process works throughout the year to provide specific users special tabulations of datasets that are not otherwise publicly available (note that "external" here relates to an entity outside the Estimates and Projections area). Each of these requests is documented, potentially approved by management, and entails an agreement with the data user regarding proper use and release of the data. Requests range in size from major survey controls like those for the American Community Survey or Current Population Survey, to relatively smaller requests such as population clock data by day or specific tabulations for research organizations collapsed into certain age categories. These data are also delivered to a variety of government agencies, such as the National Center for Health Statistics, the Bureau of Labor Statistics, the Selective Service System, and the Center for Medicare and Medicaid Services, along with state and local government agencies. They are integral to the operation of the federal statistical system.