Weissman Center for International Business
Zicklin School of Business, Baruch College/CUNY

Baruch COLLEGE ZICKLIN SCHOOL OF BUSINESS | Weissman Center for International Business | CUNY

No. 20 | WCIB Occasional Paper Series | February 2020

Exploring U.S. Census Datasets

A Summary of Surveys and Sources

Frank Donnelly[1]

Introduction

In the spring of 2020, Americans will participate in an exercise that has taken place every ten years since 1790: a count of the nation's population. The decennial census provides detailed population data that is used to reapportion seats in Congress and redraw congressional districts. Beyond this ten year count, the U.S. Census Bureau produces dozens of different datasets on an on-going basis that provide essential and geographically detailed information to policy makers, researchers, and the general public. Census data is used to allocate $880 billion federal dollars to state and local governments each year, to support a variety of programs (Reamer 2018).

In this paper, we will take a brief tour of some of the most important and widely used census datasets, to identify their distinguishing characteristics and suggest which ones to use for a given purpose. While they are generated using different methodologies and contain different sets of statistics, all of the census datasets employ similar terminology and definitions used to summarize people, housing units, and businesses into distinct population groups and geographic areas. Census data is published for a variety of legal (states, counties, cities and towns[2]) and statistical (census tracts, block groups, blocks) geographic areas. All of the data is free and in the public domain, and can be accessed from a number of sources. Following an overview of the principle datasets: the decennial census, the American Community Survey (ACS), the Population Estimates Program (PEP), and the Business Patterns and Economic Census, we will summarize some of the different sources for accessing data.

Much of this material is drawn from my new book, *Exploring the U.S. Census: Your Guide to America's Data,* published by SAGE Publications (Donnelly 2020). This researcher's guidebook covers all the main principles, fundamental datasets, and primary tasks for understanding and accessing this vast array of demographic and socio-economic data, illustrated with many examples and hands-on exercises.

---

[1] Frank Donnelly is the Geospatial Data Librarian at Baruch College, City University of New York, and an affiliate of the CUNY Institute for Demographic Research.

[2] In census terminology, towns and cities are referred to as places, which include incorporated places (legal areas) and census-designated places (concentrated population settlements that have no legal boundaries). A separate census geography called county subdivisions includes all municipalities (legal areas) for states that have them, and census county divisions (statistical areas created by the Census Bureau) for states that don't.

Datasets

Decennial Census

The decennial census is the original census dataset. It provides the foundation for many of the other datasets that the Census Bureau produces, and is a cornerstone of the federal statistical system that's used to inform policy and decision making (Anderson 2010). It is a 100% count of the U.S. population, as stipulated by Article I Section 2 of the Constitution for providing detailed population counts for reapportioning seats in Congress between the states.

The contemporary decennial census (from the year 2010 forward) captures just basic demographic information about the population and housing units: sex, age, race, Hispanic ethnicity, household and family relationships, occupied and vacant housing units, and owner and renter-occupied housing units. Each household receives a form, as does every group quarters facility (facilities where unrelated people live in a common area: penitentiaries, nursing homes, college dorms, military barracks, and a few others). There are also provisions for counting hard to reach groups such as the homeless.

Every person must be counted, and the Census Bureau follows up with non-responders and employs different methods for estimating the small percentage of the population that fails to respond. The census has been conducted by mail since 1970, but in 2020 the option to submit it on-line will be available for the first time. Under federal law, individual responses to census questionnaires are confidential and not published or disseminated for 72 years. The data is only collected and used for statistical purposes, and is summarized and published for different population groups and census geographies (US Census Bureau 2009).

The initial population count for each of the states is published by the end of the census year, and the first geographically detailed data is released for a small number of variables by March of the following year in the Public Redistricting Files, which are used for redrawing Congressional Districts. The full range of decennial census data is released on a rolling basis the following summer, within a package of files called Summary File 1. Data in this package is published in a series of tables, subdivided into population and housing unit groups and cross-tabulated by one or more variables. For a broad sample of what's included in the decennial census, the demographic profile table (DP-1) contains a cross section of the most commonly sought census variables.

Most of the decennial census is published down to the census block level, the smallest geography for which any census data is published. Figure 1 illustrates block-level data for the neighborhoods adjacent to Central Park in Manhattan.

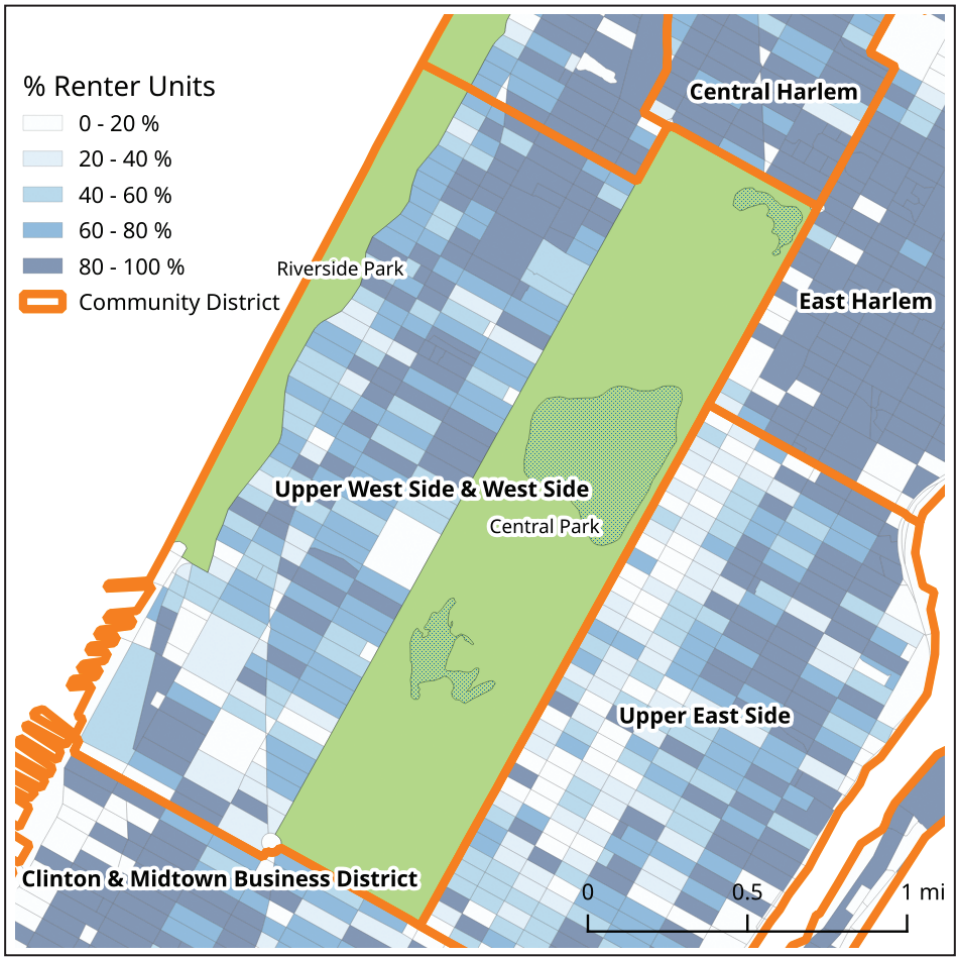

Figure 1: Percentage of Occupied Housing Units that are Renters by Census Block, Central Park Area, Manhattan NYC
Source: 2010 Census Summary File 1

When should you use decennial census data, relative to other datasets? When you need:

- Precise, 100% counts of the population, particularly for small population groups
- Reliable data published for the smallest census geographies (census blocks, block groups, and tracts)
- Just basic demographic variables
- To make historical comparisons with earlier decennial census data

From the year 2000 back, the decennial census included additional summary files with detailed socio-economic characteristics of the population that were captured on a longer sample form sent to one in six households. The American Community Survey has since replaced it, so if you are looking for anything beyond the basics you need to consult the ACS.

American Community Survey

The ACS is a rolling sample survey, designed to provide socio-economic and demographic characteristics of the population on an on-going basis. Approximately 292,000 addresses (households and group quarters) are included in the survey each month, for a total of 3.5 million addresses each year. The Census Bureau uses this sample data to create weighted estimates for hundreds of characteristics. Annual estimates are published for every geographic area in the country that has at least 65,000 people. Since annual estimates for areas smaller than this would be unreliable, the Census Bureau publishes a five-year estimate for all geographic areas above and below the 65k threshold down to the block group level[3]. Each year, new estimates are calculated by dropping the oldest year in the sample and adding the most recent year.

All ACS statistics are published as estimates with margins of error at a 90% confidence level. For example, according to the 2018 ACS, the population of Albany, NY was 97,273 +/- 35. This means that we are 90% confident that the population is somewhere between 97,238 and 97,308 people, and there is a 10% chance that the actual population falls outside this range. The smaller the population group or geographic area, the higher the margin of error will be. For instance, the population under age 18 in Albany in 2018 was 16,996, +/- 1,777.

ACS estimates must be considered as fuzzy intervals that can be used for generally characterizing an area, and should never be interpreted as exact counts (Spielman, Folch, and Nagle 2014). This fuzziness is even greater for areas with less than 65k people, as we can only characterize a 5-year time period. For example, estimates for Rensselaer, NY are only published for a 5-year period because the population of the city falls below the 65k threshold. We can say that the population of the city was approximately 9,290 +/- 20 between the years 2014 and 2018. If we wanted to make comparisons between Rensselaer and Albany, we would need to use data from the 5-year ACS for *both* places, since data is unavailable for Rensselaer in the 1-year series. The Census Bureau provides a number of recommendations for working with ACS data in a series of guidebooks (US Census Bureau 2018).

Despite these challenges, the ACS provides two principle benefits over the decennial census: it is updated on an annual basis, and it includes a much broader array of characteristics. Everything that's collected in the decennial census is included in the ACS, plus: employment, marital status, educational enrollment and attainment, veteran status, income, poverty, place of origin, housing value and rent, and much more. Like the decennial census, ACS data is published in a series of tables, and there are four profile tables that provide a sample of what's included for social (DP02), economic (DP03), housing (DP04), and demographic (DP05) characteristics. Table 1 shows the subjects included in the 2018 ACS and their associated table prefix codes, which are ID codes used for organizing and labeling the detailed tables.

Use the ACS when you:

- Need detailed socio-economic indicators not available in the decennial census
- Need the most recent data for these indicators
- Do not need precise counts, but can accept estimates with margins of error

[3] ACS data at the block group level is highly unreliable. For most applications, census tracts are the smallest feasible geography.

Table 1: 2018 American Community Survey Table Prefix Codes and Subjects

ID	Subject	ID	Subject
00	Unweighted Count (of the Sample)	15	Educational Attainment
01	Age and Sex	16	Language Spoken at Home
02	Race	17	Poverty Status
03	Hispanic or Latino Origin	18	Disability Status
04	Ancestry	19	Income
05	Citizenship Status and Year of Entry	20	Earnings
06	Place of Birth	21	Veteran Status
07	Migration and Residence 1 Year Ago	22	Food Stamps / SNAP
08	Commuting and Place of Work	23	Employment and Work Status
09	Relationship to Householder	24	Industry, Occupation, Class of Worker
10	Grandparents and Grandchildren	25	Housing Characteristics
11	Household and Family Type	26	Group Quarters
12	Marital Status and History	27	Health Insurance Coverage
13	Fertility	28	Computer and Internet Use
14	School Enrollment	29	Citizen Voting-Age Population

Population Estimates

While the decennial census is a count and the ACS is a sample survey, data from the Population Estimates Program (PEP) are annual estimates generated from a series of calculations. Using the decennial census as a base, various components (births, deaths, and migration) for different cohorts of the population (by age, sex, race, Hispanic origin) are used to create new estimates for successive years. Using this cohort component methodology, PEP data is created for the nation, states, counties, and metropolitan areas. Alternate methods are used to estimate the population for places (cities and towns) based on administrative records that track the construction and demolition of housing units.

The Census Bureau releases a new iteration of PEP data each year, which is referred to as a vintage. Each vintage includes a new year of estimates plus estimates for each year back to the last decennial census. A new

vintage completely replaces the preceding one, as estimates created for earlier years in the decade may be revised based on new information or adjustments in methodology. Once a new decennial census is conducted, the estimates for the preceding decade are revised one final time; this last series in the decade are called intercensal estimates.

Compared to the decennial census and the ACS, the PEP is a much smaller dataset that consists of just the basic variables that represent the cohorts and components of the population for large geographic areas. The components of change: births, deaths, net domestic migration, and net foreign migration, are unique to the PEP and are not published in the other two series. Given its small size, the PEP series is a simpler dataset to access and work with. Unlike the ACS, these estimates can be treated as counts and it is much easier to study the data as a time series. Figure 2 illustrates population and net migration change for Saratoga County, NY, using PEP data and an online charting application from the Missouri Census Data Center (MCDC).

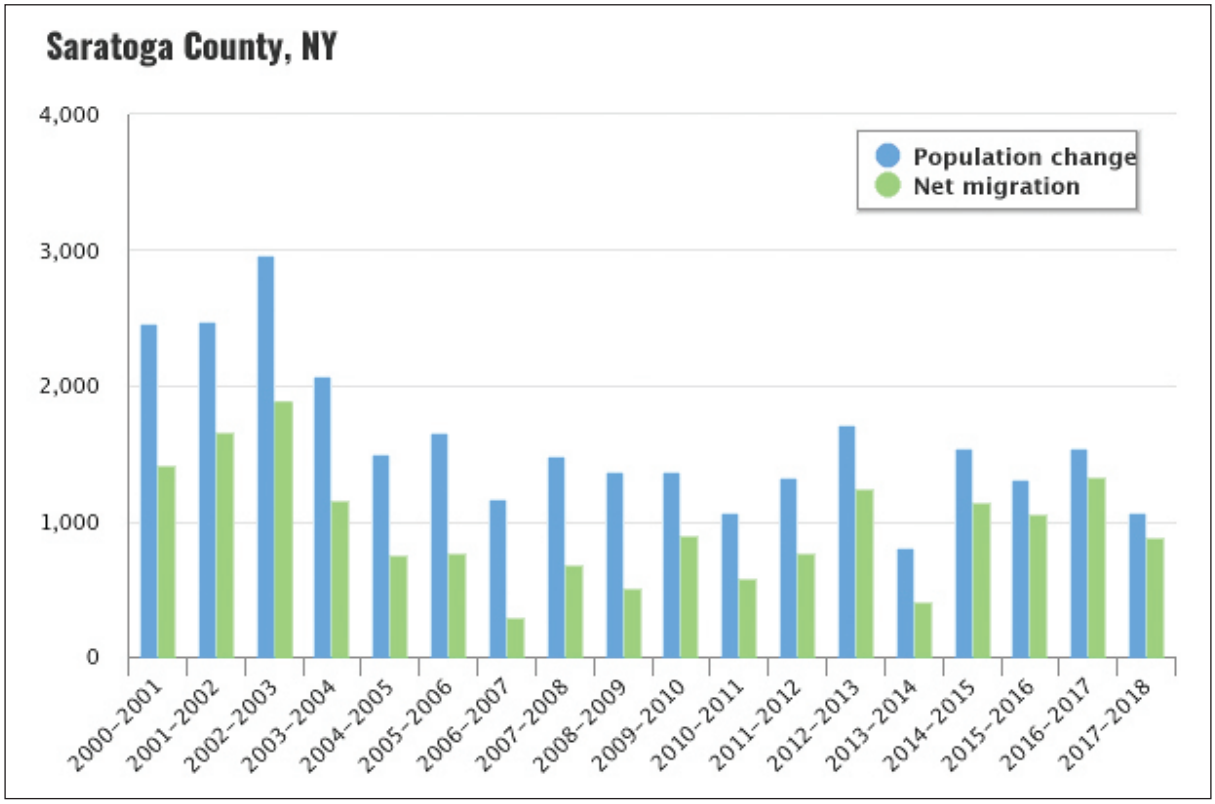

Figure 2: Population and Net Migration Change in Saratoga County, NY
Source: Missouri Census Data Center (2020) State/County Annual Population Change dataset application, http://census.missouri.edu/population/?c=36091.

Use the PEP when you:

- Want basic characteristics and annual counts of the population
- Are working with large geographic areas
- Are interested in studying components of population change

Business Datasets

The Census Bureau has been collecting data on businesses almost as long as it has been counting the nation's population (Micarelli 1998). The two datasets that we will mention here are the Business Patterns and the Economic Census. Both of these datasets count business establishments, which are defined as single physical locations where business is conducted or where services or industrial operations are performed. Establishments are assigned to industries, which are groups of businesses that produce similar products or provide similar services, using the North American Industrial Classification System (NAICS). This system is used for classifying businesses into broad groups and more specific divisions and subdivisions, with two to six digit codes that indicate related groupings and the level of detail. Table 2 illustrates the different levels of NAICS for the Scheduled Air Transportation Industry, with data from the 2017 Business Patterns for New York State. This industry includes establishments primarily engaged in providing air transportation of passengers and cargo over regular routes and on regular schedules. You can browse and search for definitions of industries at https://www.census.gov/eos/www/naics/.

.

Table 2: NAICS for the Scheduled Air Transportation Industry in New York State

NAICS Code, Title, and Level	Establishments	Employment
48-49 Transportation and Warehousing [sector]	13,107	248,505
- 481 Air Transportation [subsector]	288	28,658
-- 4811 Scheduled Air Trans. [industry group]	184	27,767
--- 48111 Sched. Air Trans. [industry]	184	27,767
---- 481111 Sched. Passenger Air Trans. [U.S industry]	158	27,364
---- 481112 Sched. Freight Air Trans. [U.S. industry]	26	403

Source: U.S. Census Bureau, 2017 County Business Patterns

The Business Patterns dataset is generated annually from the Business Register, a large administrative database updated by several federal agencies that contains every business establishment in the U.S. with paid employees. This dataset is often referred to as the County Business Patterns and ZIP Code Business Patterns, but it also includes states, metropolitan areas, and Congressional Districts. The series is relatively small and contains summaries of business establishments, employees, and payroll by geographic area and NAICS codes.

The Economic Census is a much larger undertaking that takes place every five years, in years ending in two and seven. It employs several methodologies to generate statistics, including total counts and sample surveys. It captures the variables included in the Business Patterns, as well as data on production and sales, all published by NAICS. Unlike the Business Patterns, it also includes data for cities and towns. Data is published by geography, but there are also special reports for individual sectors of the economy.

Like the population datasets, the business data is collected solely for statistical purposes and in summary form. The Census Bureau employs a variety of tactics to protect the confidentiality of businesses, to prevent people from using the summary data to obtain information on an individual business. For geographic areas or industry groups that have a small number of establishments, the Census Bureau may inject noise into the data (changing a value by a specific percentage), or may publish a footnote that indicates a range of values instead of a precise value, or may not disclose the information at all. This can make working with this data challenging, as the sum of smaller parts may not equal the whole. Data for some industries are not included but are published in different series, such as the Census of Governments, the Census of Agriculture, and the Nonemployer Statistics (for the self-employed population).

Use the Business Patterns when you:

- Want annual data on establishments, employment, and wages by industry for large geographies
- Want business data for ZIP Codes and Congressional Districts not published elsewhere

Use the Economic Census when you:

- Are doing long-term research, where timeliness is less important
- Need data on production and sales for certain industries
- Need data for ZIP Codes, towns and cities not collected elsewhere

Sources

Census data is available from a number of sources. Some websites make it easy to access profiles, which are summaries that contain a variety of statistics for a single place. Other websites provide you with options for obtaining a large amount of data for creating comparison tables, where the focus is on obtaining fewer variables for many places.

Census data is packaged in tables, which contain a selection of variables related to a certain topic. When you are searching across these various sources, you are searching across these tables and need to select one or more that contain the data you're seeking. The more basic sources will draw data from the one decennial and four ACS data profiles previously mentioned. Explore the following sources and use the ones that best meet your needs and preferences. *Exploring the U.S. Census* (Donnelly 2020) includes detailed exercises for using these resources, and each source has its own collection of tutorials and videos available on its website.

data.census.gov https://data.census.gov/

The Census Bureau's new data discovery tool, data.census.gov, has replaced the American Factfinder. You can use it to do a basic search for a geography or topic, and then casually browse around to find what you need. Using this tool to find profiles is particularly straightforward, and there are many useful charts and budding capabilities for creating thematic maps. On the other hand, keyword searching across the entire spectrum of census datasets and tables is not the best course of action. A better approach for obtaining comparison tables is to use the Advanced Search, and filter by year, dataset, geography, and topic to narrow the possibilities down to a reasonable number of tables that you can browse, select, and download.

Missouri Census Data Center https://census.missouri.edu/

The MCDC has created several applications that make it easy to assemble profile tables from the decennial census, ACS, and PEP for the entire nation. These are available from a menu on the right-hand side of their homepage. You can create tables to compare up to four geographies at once for a particular dataset, and there are applications for comparing data across recent points in time. The profiles include several charts that you can download separately as images. The MCDC also provides a number of tools for advanced users, for creating individual extracts variable by variable (Dexter) and cross walking census geographies (Geocorr).

Census Reporter https://censusreporter.org/

Originally designed for journalists, the Census Reporter makes it simple to obtain the latest estimates from the ACS. The user interface makes it easy to search by geography or topic to get data profiles and comparison tables, and there is less data to wade through as the focus is solely on the most recent ACS. The portal includes several sharp infographics, some basic thematic mapping capabilities, and accessible documentation for understanding the data.

National Historical Geographic Information System https://www.nhgis.org/

Established as part of the IPUMS project at the Minnesota Population Center, the NHGIS is the place to go if you need historic census data, back to the very first census (as opposed to most other sources, which concentrate on the present and recent past). Some users also prefer using the NHGIS for downloading contemporary data, particularly in bulk. You must register to use the site, but it's free and non-commercial.

State and Local Governments

Many state, county, and municipal governments make it easy to access census data by providing subsets of the census just for their own areas. In some cases, they may take census geographies and aggregate them to approximate areas of local interest such as neighborhoods, council districts, or regions. Some sites provide basic spreadsheet compilations, while others are sophisticated applications for exploring, mapping, and generating extracts. The most common publishers would be state data centers, local and regional planning agencies, economic development offices, departments of labor and health, and IT departments.

Census FTP and API https://www2.census.gov/ https://www.census.gov/developers/

The sources discussed so far are web portals, where users visit and navigate a website to generate extracts of data to download. There are some alternatives for more advanced users. You can visit the Census Bureau's File Transfer Protocol (FTP) site to acquire data in bulk. The site resembles a file system on a computer, where you can browse through folders to select and download an entire census dataset for a given year on a state by state basis, which you can then load into a relational database or statistical package. Alternatively, for users who want to surgically create extracts on a variable by variable basis and pull them directly into a script, you can register to get a key to access the Census Bureau's Application Programming Interface (API) and use languages like Python or R to directly access data.

References

Anderson, Margo J. 2010. "The Census and the Federal Statistical System: Historical Perspectives." *The ANNALS of the American Academy of Political and Social Science* 631: 152–62.

Donnelly, Frank. 2020. *Exploring the U.S. Census: Your Guide to America's Data*. Los Angeles: Sage Publications.

Micarelli, William F. 1998. "Evolution of the United States Economic Censuses: The Nineteenth and Twentieth Centuries." *Government Information Quarterly* 15 (3): 335–77.

Reamer, Andrew. 2018. "Census-Derived Datasets Used to Distribute Federal Funds." Counting for dollars 2020: report no. 4. Washington: The George Washington Institute of Public Policy. https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds

Spielman, Seth E., David Folch, and Nicholas Nagle. 2014. "Patterns and Causes of Uncertainty in the American Community Survey." *Applied Geography* 46: 147–57.

US Census Bureau. 2009. "Events in the Chronological Development of Privacy and Confidentiality at the US Census Bureau." Washington: US Census Bureau. https://www.census.gov/history/www/reference/privacy_confidentiality/privacy_and_confidentiality_2.html

US Census Bureau. 2018. "Understanding and Using American Community Survey Data: What All Data Users Need to Know." Washington: US Census Bureau. https://www.census.gov/programs-surveys/acs/guidance/handbooks/general.html

Weissman Center for International Business, Director:
Terrence F. Martell, Ph.D.
Saxe Distinguished Professor of Finance

Project Editor:
Lene Skou
Weissman Center Deputy Director

Design and Layout:
Rachael Cronin

For more information about this report contact the

Weissman Center for International Business
Zicklin School of Business, Baruch College/CUNY
(646) 312-2070

To access a comprehensive compilation of
information about New York City, visit NYCdata at
www.baruch.cuny.edu/nycdata

Baruch COLLEGE
ZICKLIN SCHOOL OF BUSINESS
Weissman Center for International Business
CUNY