The United States Bureau of the Census (hereafter Census Bureau) is a federal agency that produces big data that is of direct value and which also provides the foundation for analyses of other big data sources. They also produce information critical for geographic information systems in the United States.

The Census Bureau is mandated by Article I, Section II of the U.S. Constitution to enumerate the population of the United States to allow the proper allocation of members of the House of Representatives to each state. This census was first held in 1790 and then every ten years afterwards. Full data from the census is released 72 years after the census was held. With careful linking across multiple censuses, researchers can track individuals such as Civil War veterans (Costa 2017) across their full lifespan, or measure demographic changes in narrowly defined geographic regions, such as marriage rates during the boll weevil epidemic of the early 1900s (Bloome 2017).

For more recent censuses, samples of microdata are available, though with steps taken to protect confidentiality (Dreschler 2012). Information from these sources as well as census microdata from 79 other countries is available in a standardized format through the Integrated Public Use Microdata Series International Partnership (Ruggles 2015).

Starting in 1940, the Census Bureau asked additional questions for a subsample of the census. The questions, known informally as “the long form,” had questions about income, occupation, education, and other socioeconomic issues. In 2006, the long form was replaced with the American Community Survey (ACS), which covered similar issues, but which was run continuously rather than once every ten years (Torrieri 2007). The ACS has advantages associated with the timeliness of the data, but some precision was lost compared to the long form (Spielman 2014, Macdonald 2006).

Both the decennial census and the ACS rely on the Master Address File (MAF), a list of all the addresses in the United States where people might live. The MAF is maintained and updated by the Census Bureau from a variety of sources, but predominantly the delivery sequence file of the United States Postal Service (Loudermilk 2009).

Data from the MAF are aggregated into contiguous geographic regions. The regions are chosen to follow, whenever possible, permanent visible features like streets, rivers, and railroads and to avoid crossing county or state lines, with the exception of regions within Indian reservations (Torrieri 1994, Chapter 10). The geographic regions defined by the Census Bureau have many advantages over other regions, such as those defined by zip codes (Krieger 2002). Shapefiles for various census regions are available for free download from the Census Bureau website.

The census block, the smallest of these regions, typically represent what would normally be considered a city block in an urban setting, though the size might be larger for suburban and rural settings. There are many census blocks with zero reported population, largely because the areas are uninhabitable or because residence is prohibited (Freeman 2017).

Census blocks are aggregated into block groups that contain roughly 600 to 3,000 people. The census block group is the smallest geographic region for which the Census Bureau provides aggregate statistics and sample microdata (Croner 1996).

Census block groups are aggregated into census tracts. Census tracts are relatively homogenous in demographics, and self-contained within county boundaries or American Indian reservations. Tracts are relatively stable over time, with merges and partitions as needed to keep the number of people in a census tract reasonably close to 4,000 (Torieri 1994, Chapter 10).

The Census Bureau also aggregates geographic regions into Metropolitan Statistical Areas (MSA), categorizes regions on an urban/rural continuum, and clusters states into broad national regions. All of these geographic regions provide a framework for many big data analyses and helps make research more uniform and replicable.

The geographic aggregation is possible because of another product that is of great value to big data applications, the Topologically Integrated Geographic Encoding and Referencing (TIGER) System (Marx 1990). The TIGER System, a database of land features like roads and rivers and administrative boundaries like county and state lines and has formed the foundation of many commercial mapping products used in big data analysis (Croner 1996). The TIGER system allows many useful characterizations of geographic regions, such as whether a region contains a highway ramp, a marker of poor neighborhood quality, (Friesthler 2016) and whether a daycare center is near a busy road (Houston 2006).

The ACS is the flagship survey of the Census Bureau and has value in and of itself, but also is important in supplementing other big data sources. The ACS is a self-report mail survey will a telephone follow up for incomplete or missing surveys. It targets roughly 300,000 households per month. Response to the ACS is mandated by law, but the Census Bureau does not enforce this mandate. The ACS releases one year summaries for large census regions, three year summaries for smaller census regions, and five year summaries for every census region down to the block group level. This release schedule represents the inevitable tradeoff between the desire for a large sample size and the desire for up-to-date information. The ACS has been used to describe health insurance coverage (Davern 2009), patterns of residential segregation (Louf 2016), and disability rates (Siordia 2015). It has also been used to supplement other big data analysis by developing neighborhood socioeconomic status covariates (Kline 2017) and obtaining the denominators needed for local prevalence estimates (Grey 2016). The National Academies Press has a detailed guide on how to use the ACS (Citro 2007) available in book form or as a free PDF download.

The Census Bureau conducts many additional surveys in connection with other federal agencies. The American Housing Survey (AHS) is a joint effort with the Department of Housing and Urban Development that surveys both occupied and vacant housing units in a nationally representative sample and a separate survey of large MSAs. The AHS conducts computer assisted interviews of roughly 47,000 housing units biennially. The AHS allows researchers to see whether mixed use development influences commuting choices (Cervero 1996) and to assess measures of the house itself (such as peeling paint) and the neighborhood (such as nearby abandoned buildings) that can correlated with health outcomes (Jacobs 2009).

The Current Population Survey, a joint effort with the Bureau of Labor Statistics, is a monthly survey of 60,000 people that provides unemployment rates for the United States as a whole and for local regions and specific demographic groups. The survey includes supplements that allows for the analysis of tobacco use (Zhu 2017), poverty (Pac 2017), food security (Jernigan 2017), and health insurance coverage (Pascale 2016).

The Consumer Expenditure Survey, also a joint effort with the Bureau of Labor Statistics, is an interview survey of major expenditure components combined with a diary study of detailed individual purchases that is integrated to provide a record of all expenditures of a family. The purchasing patterns for the basis for the market basket of goods used in computation of a measure of inflation, the Consumer Price Index. Individual level data from this survey allows for detailed analysis of purchasing habits, such as expenditures in tobacco consuming households (Rogers 2017) and food expenditures of different ethnic groups (Ryabov 2016).

The National Crime Victimization Survey, a joint effort with the Bureau of Justice Statistics, is a self-report survey of 160,000 households per year on nonfatal personal crimes and household property crimes. The survey has supplements for school violence (Musu-Gillette 2016) and stalking (Menard 2016).

While the Census Bureau conducts its own big data analyses, it also provides a wealth of information to anyone interested in conducting large scale nationally representative analyses. Statistics within the geographic regions defined by the Census Bureau serve as the underpinnings of analyses of many other big data sources. Finally, the Census Bureau provides free geographic information system resources through their TIGER files.

Bibliography

Bloome D, Feigenbaum J, Muller C. Tenancy, Marriage, and the Boll Weevil Infestation, 1892-1930. Demography (2017 Jun.); 54(3): 1029–1049.

Cervero R. Mixed Land-Uses and Commuting: Evidence from the American Housing Survey. Transportation Research Part A: Policy and Practice (1996); 30(5): 361–377.

Citro CF, Kalton G (eds) Using the American Community Survey: Benefits and Challenges. (2007) Washington DC: The National Academies Press.

Costa DL, DeSomer H, Hanss E, Roudiez C, Wilson SE, Yetter N. Union Army Veterans, All Grown Up. Hist Methods (2017); 50: 79–95.

Croner CM, Sperling J, Broome FR. Geographic Information Systems (GIS): New Perspectives in Understanding Human Health and Environmental Relationships. Stat Med (1996); 15: 1961-1977.

Davern M, Quinn BC, Kenney GM, Blewett LA. The American Community Survey and health insurance coverage estimates: possibilities and challenges for health policy researchers. Health Serv Res. (2009 Apr.); 44(2 Pt 1): 593-605.

Dreschler J, Reiter JP. Sampling With Synthesis: A New Approach for Releasing Public Use Census Microdata. J Amer Statist Assoc (2012); 105(492): 1347-1357.

Freeman NM. Nobody lives here: The nearly 5 million Census Blocks with zero population (2014 Apr. 15). <http://tumblr.mapsbynik.com/post/82791188950/nobody-lives-here-the-nearly-5-million-census>. Accessed August 6, 2017.

Freisthler B, Ponicki WR, Gaidus A, Gruenewald PJ. A Micro-Temporal Geospatial Analysis of Medical Marijuana Dispensaries and Crime in Long Beach California. Addiction (2016 Jun.): 111(6): 1027–35.

Grey JA, Bernstein KT, Sullivan PS, Purcell DW, Chesson HW, Gift TL, Rosenberg ES. Estimating the Population Sizes of Men Who Have Sex With Men in US States and Counties Using Data From the American Community Survey. JMIR Public Health Surveill. (2016 Apr. 21); 2(1): e14.

Houston D, Ong PM, Wu J, Winer A. Proximity of licensed childcare to near-roadway vehicle pollution. Am J Public Health (2006); 96(9): 1611–1617.

Jacobs D, Wilson J, Dixon SL, Smith J, Evens A. The Relationship of Housing and Population Health: A 30-Year Retrospective Analysis. Environmental Health Perspectives (2009); 117(4): 597–604.

Jernigan VBB, Huyser KR, Valdes J, Simonds VW. Food Insecurity among American Indians and Alaska Natives: A National Profile using the Current Population Survey-Food Security Supplement. J Hunger Environ Nutr (2017); 12(1): 1-10.

Kline K, Hadler JL, Yousey-Hindes K, Niccolai L, Kirley PD, Miller L, Anderson EJ, Monroe ML, Bohm SR, Lynfield R, Bargsten M, Zansky SM, Lung K, Thomas AR, Brady D, Schaffner W, Reed G, Garg S. Impact of pregnancy on observed sex disparities among adults hospitalized with laboratory-confirmed influenza, FluSurv-NET, 2010-2012. Influenza Other Respir Viruses. (2017 Jul. 13).

Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas--the Public Health Disparities Geocoding Project. Am J Public Health (2002 Jul.); 92(7): 1100-2.

Loudermilk CL, Li M. A National Evaluation of Coverage for a Sampling Frame Based on the Master Address File. Proceedings of the Joint Statistical Meeting (2009).

Louf R, Barthelemy M. Patterns of Residential Segregation. PLoS One (2016 Jun 17); 11(6): e0157476.

Macdonald H (2006) The American Community Survey: Warmer (More Current), but Fuzzier (Less Precise) than the Decennial Census, J Amer Plan Assoc (2006); 72(4): 491-503.

Marx RW. The Census Bureau's TIGER system. N Z Cartogr Geogr Inf Sys. (1990 Jan); 17(1): 17-113.

Menard KS, Cox AK. Stalking Victimization, Labeling, and Reporting: Findings From the NCVS Stalking Victimization Supplement. Violence Against Women (2016 May); 22(6): 671-91.

Musu-Gillette L, Zhang A, Wang K, Zhang J, Oudekerk BA. Indicators of School Crime and Safety: 2016. <https://www.bjs.gov/content/pub/pdf/iscs16.pdf>. Accessed August 6, 2017.

Pac J, Waldfogel J, Wimer C. Poverty among Foster Children: Estimates Using the Supplemental Poverty Measure. Soc Serv Rev (2017 Mar.); 91(1): 8-40.

Pascale J, Boudreaux M, King R. Understanding the New Current Population Survey Health Insurance Questions. Health Serv Res (2016 Feb.); 51(1): 240-61.

Rogers ES, Dave DM, Pozen A, Fahs M, Gallo WT. Tobacco cessation and household spending on non-tobacco goods: results from the US Consumer Expenditure Surveys. Tob Control (2017 Mar. 16); pii: tobaccocontrol-2016-053424.

Ruggles S, McCaa R, Sobek M, Cleveland L. The IPUMS Collaboration: Integrating and Disseminating the World’s Population Microdata. J Demogr Economics (2015 Jun.); 81(2): 203-216.

Ryabov I. Examining the role of residential segregation in explaining racial/ethnic gaps in spending on fruit and vegetables. Appetite (2016 Mar. 1); 98: 74-9.

Siordia C. Disability Estimates between Same- and Different-Sex Couples: Microdata from the American Community Survey (2009-2011). Sex Disabil (2015 Mar.); 33(1): 107-121.

Spielman SE, Folch D, Nagle N. Patterns and causes of uncertainty in the American Community Survey. Appl Geogr (2014 Jan.); 46: 147–157.

Torrieri NK. Geographic Areas Reference Manual. 1994 (Nov.) <https://www.census.gov/geo/reference/garm.html>. Accessed August 7, 2017.

Torrieri N. America is Changing, and So Is the Census: The American Community Survey. The American Statistician (2007 Feb.); 61(1): 16-21.

Zhu SH, ZhuangYL, Wong S, Cummins SE, Tedeschi GJ E-cigarette use and associated changes in population smoking cessation: evidence from US current population surveys. BMJ (Clinical Research Ed.) (2017 Jul. 26); 358: j3262.