Chapter 3
Data Sources for Demography

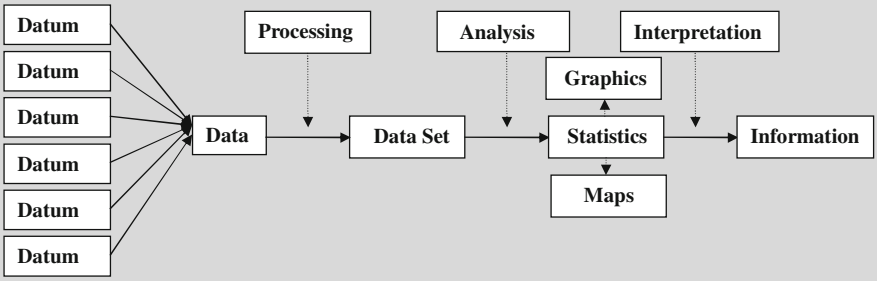
Abstract While all aspects of demography rely on data, applied demography employs a wider range of demographic data and data from other fields than do traditional demographic analyses. Understanding the nature, attributes and availability of relevant data is critical for the effective application of data to real-world problems. This chapter provides an overview of the types of data used by demographers and the sources of these data.

3.1 Introduction

The grist for the demographic mill is data, and virtually every demographic activity involves the examination of statistics for the population or topic under consideration. The purpose of this chapter is to identify and describe the broad range of data sets of use to demographers as well as others who are interested in population characteristics. While it is impossible to consider all possible types of demographic data and their sources, this article addresses the most important data resources available to demographers, the means through which the data are generated, and the ways in which the data can be accessed. Exhibit 3.1 describes the process through which raw data is converted into information for use by applied demographers.

R. K. Thomas, *Concepts, Methods and Practical Applications in Applied Demography*, https://doi.org/10.1007/978-3-319-65439-3_3

Exhibit 3.1: The Data Conversion Process



3.2 Demographic Data

Data represents the fuel for the demographic engine and virtually all demographic analyses are data heavy. An applied demographer must be familiar with various types of data, their characteristics and their sources and be something of a statistician. The following sections introduce the student to this important component of the demographic endeavor.

3.2.1 Data Types

Data utilized by demographers can take a variety of forms that include the following:

Raw data. Raw data represents the quantifiable attributes associated with any individual or group—e.g., the age of every student in a demography class. Each "record" (e.g., person, group) will have a series of attributes associated with it. Raw data by themselves are not very useful and need to be processed into "information" (as illustrated in the graphic above). Demographers conducting primary research generate raw data.

Statistics. Statistics are raw data that have been processed and subjected to some level of statistical analysis (that is, converted into information). Descriptive statistics provide a basic picture of the population (e.g., the age range and median age of the demography students) while inferential statistics examine relationships among various attributes of the population under study (e.g., the relationship between the age of students and political orientation).

Data sets. A data set is a collection of data in either raw or summary form. All of the attributes for all subjects (in raw data form) is one type of data set. The summarized data based on the raw data represents another type of data set (summary data). Data sets are typically utilized for analysis purposes when the demographer does not have access to the raw data.

Reports/Tables. Data presented in report or table form essentially represents a set of statistics. A table comparing the median age of students in all classes in the university would be presented in table form. Most secondary data is presented in table or report format.

Graphics/maps. Graphics (charts, graphs) and maps represent methods for displaying data although they may serve as the basis for analysis. Charts and graphs typically depict summary data (e.g., the proportion of students of each age in the class) but may depict the raw data (e.g., a scattergram of all ages). Maps depict the spatial (geographic) distribution of one or more attributes. Maps can typically display more information than charts and graphs and, by virtue of their link to geography, lend themselves to an examination of relationships between variables (e.g., the location of fast food restaurants relative to the distribution of the low-income population).

3.2.2 Using Demographic Data

There are a variety of ways in which demographers use data in applying their knowledge to the solution of real-world problems. It might be useful to think of this process as following a logical sequence from the initial enumeration of the population to the ultimate application of knowledge.

Counting. The first step in any demographic analysis involves counting the number of people within the population under study. The count is the first demographic fact and provides the basis for calculating a variety of other attributes.

Describing. Counting the population by itself does not provide adequate information for understanding the nature of that population. Data must be collected that describe that population in terms of its salient attributes (e.g., age distribution, racial makeup, income level).

Comparing. Numbers by themselves do not mean much unless you have some standard for comparison. Is the size of the population (e.g., 10,000 people) large or small? That depends on the standard for comparison. It is typical to compare figures for one population to another population (e.g., County A with County B) and/or compare the current population figure with a previous time period. These comparisons allow an assessment of the figure in question.

Explaining. The objective of the above steps is to allow the analyst to explain demographic phenomena that are observed within the environment under study. This is the point at which relationships between variables are identified and quantified. Thus, we can explain the declining fertility rate among the non-Hispanic white population by examining the fertility rate in relation to the population's median age. In short, non-Hispanic whites in the childbearing ages are aging out—that is, growing too old to have children.

Understanding. The identification of relationships provides the basis for interpreting the observed phenomena and developing an understanding of what is transpiring. This produces the "ah ha" moment when the meaning behind the

statistics is revealed to the analyst. Based on the example above, we now understand not only *why* there are fewer births among non-Hispanic whites but why the non-Hispanic white population is shrinking as a proportion of the total population.

Predicting. The goal of any science is to be able to predict future developments and/or conditions. Much of the effort of applied demographers is directed toward predicting the size and characteristics of the population in the future. Demographers have a built-in advantage in this regard in that they can use cohort analysis to extrapolate into the future. If we know, for example, how many 55-year-olds there are in the U.S. population, we can predict how many people will be enrolling in Medicare ten years from now (minus a known attrition rate). Not only that, but we will know the characteristics of those enrolling in Medicare at that point.

Addressing. Finally, we get to the most applied aspect of applied demography. Based on all of the above steps in the analysis, what do we do with what we have found out? The knowledge that has been developed to this point can be disseminated in various ways to policy setters, lawmakers, and decision makers. It can be used to make recommendations to government agencies, private organizations and businesses. It can provide the basis for the development of new programs or the modification or elimination of existing initiatives. This is the point where the knowledge is applied directly to the resolution of real-world problems.

3.2.3 Data Timelines

While the first thought when profiling a population is to identify the current population attributes, demographers actually may consider three time periods when doing their analyses.

Past. Having data for past time periods allows the analyst to identify historical trends for the population in question. This supports comparisons over time and provides the basis for projections into the future. In demography, the past is typically (but not always) a valid basis for predicting the future.

Present. For most purposes, analysts determine the current demographic profile of a population. This is typically where the analysis begins and then works backward or forward depending on the objectives of the analysis. This "snapshot" of the current situation provides the basis for subsequent analyses.

Future. Given the importance of prediction in applied demography, most projects focus on the future. The goal is to formulate a picture of the future based on past and present data and assumptions about future developments. This activity may be as simple as predicting the number of first-graders in a school district five years from now or as complex as projecting the impact of global warming on the population of sub-Saharan Africa. The process is the same in any case, extrapolating past trends into the future.

3.3 The Importance of Geography

Virtually all demographic phenomena are linked to geography, and demographics data are invariably presented for some geographic unit (e.g., nation, state, city, census tract). This is particularly important given the fact stated above that virtually no phenomena (especially social phenomena) are randomly distributed. As noted, spatial relationships represent important information and provide the basis for much demographic and epidemiologic analysis. Geography defines the context in which demographic actors play out their roles.

A knowledge of geography is also important for demographers since different types of data are available for different geographies. This has significance for determining the geographic level at which an analysis will take place. Further, some geographic boundaries change over time (e.g., ZIP Codes, city limits, political districts), and these changes often have important implications for the reported data. (More detail on the geographic units utilized by demographers is presented in Chap. 4).

3.4 Data Collection Methods

The methods of data collection discussed below are divided into four general categories: censuses, registration systems, surveys, and synthetically-produced data. Censuses, registries, and surveys are the more traditional sources of demographic data, although synthetically produced statistics such as population estimates and projections have become standard tools for most planning, marketing and business development activities.

3.5 Census

A census involves a complete count of individuals (or entities) residing in a specific place at a specific time. The U.S. Census Bureau (within the Department of Commerce) has conducted population censuses since 1790. Subsequently expanded to the census of population and housing, it has been conducted every 10 years (in years that end in zero) since. Originally established as a means of determining the number of residents in each state for the purpose of allocating Congressional seats, the content of the census has expanded over time to include hundreds of additional data elements.

The expanded census was conducted through 2000 and included a short form administered to every household and a long form administered to one in six households. For the 2010 census the long form was discontinued and only the short form administered. The American Community Survey (ACS), described below, was

instituted to replace the long form, with data collected via the ACS in 2010 used to supplement data collected via the short form. Exhibit 3.2 provides an overview of the 2010 census and discusses some of the issues surrounding its administration.

In order to assure a complete count of residents at the time of the census, the street address is used as the key locator. Using data provided by the U.S. Postal Service, the Census Bureau identifies every postal address in the country and uses this information as the basis for collecting data. For the majority of the population with a single permanent address this approach is effective, and most census forms today are returned by mail. However, there are exceptions that require additional effort on the part of the Census Bureau, the canvassing of migrant workers and homeless individuals for example. For those addresses for which no census form is returned, in-person interviews at the place of residence may be required.

By definition, a census includes a complete count of the population. However, it is increasingly difficult to strictly apply this term to the U.S. decennial census. While the census ostensibly counts every resident, it falls short of a true census in two respects. First, every decade a certain segment of the population is missed in the enumeration resulting in some level of undercount. While the undercount is typically less than 3%, its mere existence creates myriad problems. This undercount tends to be concentrated among certain segments of the population, resulting in overrepresentation of some groups and underrepresentation of others. This fact has important implications, since the results of the census are used as the basis for reallocating Congressional seats and allocating government funds. Because of the undercount, the initial release of census data every ten years produces a spate of lawsuits questioning the accuracy of the census. The controversy surrounding the undercount has only become more intense in view of the large number of undocumented residents entering the U.S. in recent years.

Exhibit 3.2: 2010 Census of Population

In 2010, the US Census Bureau administered the twenty-third decennial census. The 2010 data collection effort involved the mailing of questionnaires to every known household. Each household received a form with the ten core questions (the short form). Prior to the mailing of the questionnaires, postcards were mailed to every household to alert residents to the coming survey instrument. Included with the questionnaire were instructions for completing the form and an envelope for return mail. In addition to the mail-in option, Americans were given the option of completing the census questionnaire on line via the Internet. Other forms of the questionnaire are used for individuals with non-household living arrangements such as those living in group quarters.

The core questions in the 2010 census related to the information required for political redistricting purposes. These questions captured data on the age, sex, race/ethnicity and tenure of each household member as well as on the relationships of household members. This information was subsequently used nationally for the allocation of seats in the U.S. House of Representatives and,

at the state level, for adjusting the boundaries of Congressional districts or drawing boundaries for new districts as appropriate. Information on the 2010 census and its implementation is available at www.census.gov.

In past censuses, a long form covering more than 50 topics was mailed to one in six households. That form was replaced for the 2010 census by the American Community Survey which collects data from a representative sample of households. (See Exhibit 3.3 for additional information on the American Community Survey).

The second factor diminishing the enumeration's value as a census is the fact that a large portion of the data on population characteristics is obtained from a random sample of the nation's households. In 2010, for example, only 10 questions were included on the short form that went to every household. The remaining data were obtained via the American Community Survey which involves a sample of 1 in 20 households. This small sample size means that detailed data are not as likely to be available for small geographic areas and that a large margin of error is generated. While the use of sampling significantly reduces the cost of taking a census and the ACS assures a steady flow of data between decennial censuses, it generates figures that some might assume (incorrectly) to represent complete counts.

The decennial census in the United States theoretically provides the most accurate enumeration of the population since it represents an attempt to include all residents. Because of certain shortcomings, however, the census count is supplemented by other sources of data today.

The decennial census today collects data on the number of persons residing in each living unit (e.g., house, duplex, apartment, and dormitory), their age, race/ethnicity and sex, and the relationship of those individuals to each other. On the ACS form, data are gathered on the age, race, ethnicity, marital status, income, occupation, education, employment status, and industry of employment for each resident. Questions related to the dwelling unit in which the respondent lives elicit information on the type of dwelling unit (e.g., apartment or duplex), ownership status, value of owned house, monthly rent, age of dwelling unit, and a number of other topics.

Statistics generated by virtue of the census are available for virtually every formally designated geographic unit in the United States. These statistics are disseminated for states, counties, zip codes, metropolitan areas, and cities. Statistics are also produced for specially designated areas created by the Census Bureau, including blocks, census tracts, block groups, and block numbering areas.

Very little of the results of the census are presented in print form today, although detailed data can be readily accessed via the Internet. The Census Bureau makes

certain databases—referred to as summary tape files (STFs)—available to those who want to further analyze the data. These databases do not include raw data (i.e., individual records) from the census but preselected aggregations of data. Public use microdata samples (PUMS) do include raw data and are available from the 2010 census, stripped of any information that would identify individual respondents. PUMS files involve a sample of records from areas containing at least 100,000 persons.

After the 1980 census, many private data vendors began to acquire census data and sell repackaged data to the public. In fact, joint public-private projects were involved in converting census data to the ZIP Code level, a geographic unit with a great deal of utility for the business community. Private sector marketing of census data was even heavier after the 1990 census, with commercial data vendors providing population estimates and projections at the census tract level during the intercensal period. Case Study 3.1 describes the use of demographic data for political redistricting.

Case Study 3.1: Redistricting after the 2010 Census: Monterey County, California, Board of Supervisors

After each decennial U.S. Census, most jurisdictions are required to evaluate whether election districts meet the population equality requirement ("one person, one vote"). This is required not only of Congressional districts, but of any local entity that elects governing boards by district, including counties, cities, school districts, water districts, and other special districts. There is much redistricting activity after each decennial Census.

If the election districts adopted after the previous Census are not balanced using the new Census counts, they must be re-drawn. In addition to equalizing total populations, redistricting plans must meet Federal Voting Rights Act (FVRA) requirements. Among other things, the FVRA requires that members of certain protected groups, including Hispanics, be given the opportunity to elect representatives of their choice.

A demographic consultant was hired in 2011 to help the Monterey County, California, Board of Supervisors with its redistricting process. To ensure a fair and transparent process with as low a risk of litigation as possible, the Board appointed a citizens advisory committee to recommend a redistricting plan.[1] The consultant provided demographic support to the committee, and

[1]Many jurisdictions convene some sort of citizens' redistricting advisory committee to make recommendations to the governing board. In California, Proposition 11, passed in 2008, assigned the drawing of California's state and congressional election districts to a citizens committee. Such committees are charged with taking legal requirements into account, often along with placing priority on easily identifiable boundary lines and compact districts. If instructed to do so, they recommend plans with only one incumbent per election district (but sometimes are instructed to disregard incumbency altogether). Representatives of protected groups on these committees, in our experience, often advocate for the creation of election districts in which their groups can easily elect candidates of choice.

after several months of deliberation, the committee recommended a plan that made very minor modifications to the 2001 districting plan. Just two voting precincts were moved from one supervisorial district to another to equalize total populations.

Despite nearly unanimous support from the citizen's advisory committee and enthusiastic support from the Board of Supervisors, the City of Salinas (with more than one-third of the county's total population) objected to the redistricting plan. The county's plan divided Salinas among four supervisorial districts, and the city's leaders wished to be divided between only two. However, there is no legal prohibition on fragmenting large cities, so the city could not legally object to the plan on that basis. Instead, attorneys were hired by the city to support a plan that kept Salinas in two districts and featured more Hispanic-majority districts. It therefore became necessary to re-evaluate compliance with the FVRA under the county's and the city's plans, and the consultant implemented that evaluation.

The FVRA requires that districting plans allow members of protected groups that have large populations and are geographically compact, like Monterey County's Hispanics, to elect representatives of their choice. Court rulings on how to test whether this requirement is met have resulted in demographers' relying on several measures of ability to elect:

- Total population shares
- Voting age population (VAP) shares
- Citizen voting age population (CVAP) shares
- Registered voter shares
- Actual voter shares.

Figure 3.1 shows Monterey County's values and estimates for each measure. Note that in this case, the Hispanic share shrinks with each, increasingly restricted, measure. While Hispanics comprised 56% of the total county population, they comprised only 33% of those eligible to vote (CVAP) and 27% of actual voters. This means that it was rather challenging to create districting plans that potentially provided Hispanics the opportunity to elect representatives of their choice. In the consultant's experience working with many Monterey County jurisdictions since 1990, local Hispanic leaders generally preferred CVAP shares of 60% or more in order to offset Hispanics' low voter registration and turnout rates.

The County's plan created two Hispanic-majority election districts, with 63 and 65% Hispanic CVAP shares. The city of Salinas' plan created three districts, each with barely 50% Hispanic CVAP shares. The county argued that the city's plan was "retrogressive"; that is, Hispanics would have less opportunity to elect representatives than they had under the plan adopted in 2001. In the end, the county's "minimum change" plan was precleared by the U.S. Department of Justice (before the U.S. Supreme Court suspended Section 5 of the FVRA) and no lawsuit was filed by the city. Based on input

Category	Total	Hispanic	Hispanic share
Population, all ages*	405,087	225,627	56%
Voting age population (VAP)*	294,083	144,987	49%
Estimated citizen voting age population (CVAP)**	211,716	70,920	33%
Registered voters Nov 2008***	145,596	45,225	31%
Actual voters Nov 2008***	118,629	32,560	27%

* Census 2010 counts
**Based on American Community Survey 2005-2009 rates
***Totals from Monterey County Registrar of Voters, estimated Hispanic counts based on Spanish Surname estimates

Fig. 3.1 Categories used to evaluate districting plans, all of Monterey County, in 2011

from the demographic consultant, a satisfactory arrangement with regard to redistricting was achieved[2].

Source Gobalet, J., and S. Lapkoff, Lapkoff & Gobalet Demographic Research, Inc., www.demographers.com.

3.6 Economic Census

In addition to counting people, the Census Bureau also counts businesses. Economic censuses can be traced back to the early nineteenth century, although it was not until 1929 that continuous data gathering for a broad range of business entities was begun. The modern economic census was initiated in 1954 and is conducted every five years (currently in years ending in 2 and 7). The census covers businesses engaged in retail trade, wholesale trade, service activities, mineral industries, transportation, construction, manufacturing, and agriculture, as well as government services. The information collected through the economic census includes data on sales, employment, and payroll, along with other, more specialized data. These data are available for a variety of geographic units, including states, metropolitan areas, counties, and incorporated places of 2500 or more population.

Every business is classified using the North American Industrial Classification system (NAIC). The assigned NAIC code allows businesses to be grouped into standard categories for statistical purposes. Aggregated information on businesses within the NAIC categories is available at the county level with distribution primarily through the Internet. For example, for a particular U.S. county the 2007 economic census found 1740 physician offices with 16,176 employees and payrolls of $2.6

[2]Shelby County vs. Holder, 570 U.S. 2 (2013).

billion, 234 chiropractic offices with 994 employees and payrolls of $33 million, and 122 medical laboratories with 1772 employees and payrolls of over $92 million.

3.7 Registration Systems

A second method of data collection that generates information for demographers is represented by registration systems. A registration system involves the systematic compilation, recording, and reporting of a set of events, institutions, or individuals. The implied characteristics of a registry include the regular and timely recording of the phenomenon in question. Most registration systems relevant to this discussion are maintained by some branch of government, although other sponsors of registration systems exist as well.

For demographers the best-known registration activities in the United States are those related to "vital events", such as births, deaths, marriages, and divorces. The most extensive registration systems are maintained by the National Center for Health Statistics and the Centers for Disease Control and Prevention (CDC). Other useful systems are maintained by the Social Security Administration (SSA), the Centers for Medicare and Medicaid Services (CMS), and Immigration and Customs Enforcement (ICE). Lists maintained on members by trade groups and professional associations (such as the American Medical Association and the American Hospital Association) are placed in this category because such lists have many of the characteristics of registries.

A variation on registration systems increasingly deployed by demographers involves administrative records. Administrative records systems are not necessarily intended to be registries of all enrollees or members of a group of events, organizations, or individuals but to provide a record of the transactions of those included with that group. Thus, the list of all Medicare beneficiaries (enrollees) would constitute a registry but the data generated by virtue of the beneficiaries encounters with the healthcare system would be considered administrative records (since not all beneficiaries would use services during a given time period).

Administrative records can serve a useful function to the extent that they provide access to sources of data not otherwise available. However, unlike other forms of data generation such as censuses and surveys, the raw data are not strictly under the control of those who establish the data file. Administrative records may be submitted by a variety of parties, creating inherent problems in data quality and standardization. A great deal of effort is currently being expended to improve the accessibility of data maintained by federal agencies. For example, Medicare data on the number of current enrollees are now available for all U.S. counties, as are year-to-year migration data from the Internal Revenue Service. The Census Bureau is exploring the use of registries and administrative records as sources of data to supplement its traditional data gathering activities.

3.7.1 Vital Statistics

As noted above, vital statistics include data on births, deaths, marriages, and divorces. The collection of vital data has a long history in the United States, predating the Declaration of Independence by many years. The collection of data on vital events is initially the responsibility of local government (i.e., city or county government). A local court clerk's office is responsible for the recording of marriages and divorces, while the local health department is the primary collector of birth and death statistics. Data collected at the local level are forwarded to the appropriate vital statistic registry within the respective state governments. The state agency compiles the data from the various localities and subsequently transfers the data (in the case of births and deaths) to the National Center for Health Statistics (NCHS). The NCHS has responsibility for compiling and publishing vital statistics for the nation and its various political subdivisions.

A standard birth certificate is used in the United States to collect data on the time and date of birth, place of occurrence and the mother's residence, birth weight, pregnancy complications, mother's pregnancy history, mother's and father's age and race/ethnicity, and mother's education and marital status. Information gathered on the standard death certificate includes age, race/ethnicity, sex, place of residence, usual occupation, and industry of the decedent, along with the location where the death took place. In addition, data are collected on the immediate and secondary causes of death, as well as on any other significant conditions. A separate certificate is used for fetal deaths. There is some variation in the content of birth and death certificates from state to state, although there are certain data elements that are always collected.

Birth and death statistics are traditionally available in government publications and increasingly electronically via the internet. The compiled statistics are typically presented for both the place of occurrence of the vital event (e.g., the location of the hospital) and the place of residence of the effected individual. Considerable detail is provided by the NCHS for a wide range of geographic units including states, metropolitan statistical areas (MSAs), counties, and urban places. Data for other geographic areas may be available through state and local governmental agencies. Yearly summary reports are produced and published by the National Center for Health Statistics, and periodic updates are available through the monthly vital statistics reports. Local and state health departments are increasingly making birth and death statistics available on line.

Marriage and divorce registration areas (MRAs and DRAs) are established using the same criteria as birth and death registration systems. Standard data collected on the marriage certificate includes age of spouses, type of ceremony (civil or religious), and previous marital status of spouses, as well as race and educational status of the bride and groom. The data available on marriages and divorces varies from state to state and, since the NCHS discontinued its marriage and divorce registries, there is no nationwide system for aggregating marriage and divorce data.

3.7.2 Immigration Data

Data on immigration patterns and the characteristics of immigrants historically have been of interest to demographers because of the implications of these phenomena for population change. Today, however, data on immigration are of increasing interest due to the growth of illegal immigration and concerns over the impact of immigration on other demographic processes and attributes. Monitoring international migration is a responsibility of the federal government, and the agency responsible for monitoring and reporting on immigration trends is the Immigration and Customs Enforcement (ICE) agency, formerly the Immigration and Naturalization Service (INS), now within the Department of Homeland Security. Data are collected related to legalization applications, refugees, asylum applicants, nonimmigrant entries, naturalizations, and enforcement activities and made available by means of published reports and the Internet. Because of the increase in illegal immigration, a growing amount of data is generated as a result of border monitoring and internal police activities. Annual estimates of illegal immigration are generated by ICE. Additional data on immigration can be obtained from www.ice.gov.

Data on visas issued is maintained by the Department of State, and data on immigrant visas are available on everyone legally entering the United States. After a person is admitted to this country, visa and adjustment forms are forwarded to the ICE data-capture facility for processing. Information collected includes port of admission, country of birth, last residence, nationality, age, sex, occupation, and the ZIP Code of the immigrant‘s intended residence.

Data on immigration are made available through yearly statistical summaries, more frequent shorter reports, and via the Internet. While the published reports contain data for states and MSAs, tabulations by county and zip code are possible by accessing ICE data files.

3.8 Surveys

A sample survey involves the administration of an interview form to a portion of a target population that has been systematically selected. The sample is designed so that the respondents are representative of the population being examined. This allows conclusions to be drawn for the total population based on the data collected from a sample.

The use of sample surveys has several advantages relative to the census and registry methods. Two of the major advantages are more frequent data collection and the ability to probe more deeply into the subject under study. The relatively small sample sizes for such surveys have the additional advantages of quicker turnaround time and easier manipulation than large-scale operations such as the census.

On the other hand, surveys have their disadvantages. Since they involve a sample, figures generated are estimates resulting in some slippage in accuracy relative to censuses. Other potential problems include interviewer bias, misinterpretation of survey items or inaccurate or dishonest responses on the part of respondents. Perhaps the most serious shortcoming related to demographic analysis is the inability to compile adequate data for small geographic units due to small sample sizes.

Much of what demographers know about the demographic attributes of a population is based on sample surveys conducted by various government agencies, research institutes and private vendors.

The federal government is the major source of survey data related to a variety of topics. Through various agencies, the federal government administers a number of ongoing surveys that involve information of interest to demographers. The National Institutes of Health and the Centers for Disease Control and Prevention conduct surveys that generate data of interest to health demographers. The Census Bureau conducts surveys related to population characteristics, housing and economic trends. The Department of Labor conducts surveys related to employment, occupational trends and so forth.

As previously noted, the American Community Survey (ACS) was introduced by the Census Bureau as a replacement for the long-form data collection instrument for the decennial census. Increasingly, demographers and others have come to rely on data from the ACS for a wide variety of uses despite the shortcomings of this survey. Exhibit 3.3 presents information on the ACS.

A number of private organizations also conduct surveys of interest to demographers. There are research institutes and "think tanks" that collect data on reproductive practices, immigrant characteristics, consumer behavior patterns and other topics. The results of these surveys are typically not available for levels of geography below the nation, although some may generate state-level data. Demographic-oriented organizations like the Population Reference Bureau may also conduct surveys, although they typically repackage data from other surveys.

Commercial data vendors also conduct surveys that contain data useful to demographers. Various surveys are conducted on consumer behavior and these invariably have information on the demographics of respondents. At least two vendors conduct national surveys annually on health-related characteristics and health behavior. Other data vendors may extract health-related data from national syndicated surveys and package this information with their demographic data. Some of these data sets are considered proprietary and generally are only available to established clients. Other data may be available for sale to the public.

Exhibit 3.3: The American Community Survey

The American Community Survey (ACS) is an ongoing survey that collects data every year on most U.S. communities. The ACS includes 69 questions on topics such as income, household expenses, employment, education, and work commutes. With full implementation in 2005, the sample included 3 million addresses throughout the U.S. and another 36,000 in Puerto Rico. In 2006, approximately 20,000 group quarters were added to the ACS database. Approximately 250,000 interviews are conducted each month with some 2.5% of the population administered the ACS in any given year.

Unlike the decennial census, the ACS involves continuous measurement of the topics under study. Continuous measurement has long been viewed as a possible alternative method for collecting detailed information on the characteristics of population and housing; however, it was not considered a practical alternative to the decennial census long form until the early 1990s. At that time, demands for current, nationally consistent data from a wide variety of users led federal government policymakers to consider the feasibility of collecting social, economic, and housing data continuously throughout the decade. The benefits of providing current data, along with the anticipated cost savings from a scaled-back census and more efficient operations, led the Census Bureau to plan the implementation of what came to be called the American Community Survey (ACS).

The following criteria were considered important for an effective on-going survey:

- Data would be collected continuously by using independent monthly samples.
- Three modes of data collection would be used: mail-out, telephone non-response follow-up, and personal visit non-response follow-up.
- The survey reference date for establishing housing unit occupancy status and for many characteristics would be the day the data were collected. Certain data items would refer to a longer reference period (for example, "last week," or "past 12 months").
- The survey's estimates would be controlled to intercensal population and housing estimates.
- All estimates would be produced by aggregating data collected in the monthly surveys over a period of time so that they would be reported annually based on the calendar year.

Data generated by the ACS are presented for various levels of census geography, with the lowest level being the census block group. Since the sample is not large enough to produce accurate estimates for all geographies in any particular year, the results of the ACS are published in three temporal versions: 1-year data, combined 3-year data, and combined 5-year data. The more years that are combined the greater the sample size and the more reliable the estimates.. For larger geographies, one year of data may suffice but more often than not a smaller community or lower level of geography will

necessitate the combining of years. Combined data, of course, have the disadvantage of representing different time periods, sometimes combining data separated by four years in time.

While the ACS does not have the statistical power of the one-in-six household long form used by the Census Bureau in the past and demographic purists raise some issues with the methodology, the benefit of having continuous data collection outweighs any drawbacks. The most direct way to access data from the American Community Survey is through the "American Factfinder" function on the Census Bureau website accessed at www.census.gov.

3.9 Synthetic Data

Synthetic data refers to statistics that are produced in the absence of actual data using models that simulate reality. Synthetic data are generated by merging existing demographic data with assumptions about population change to produce estimates, projections, and forecasts. These data are particularly valuable given that census and survey activities are constricted because of budgeting and time considerations. Further, there are situations in which no actual data are available for a particular population, geographic unit or time period. Consequently, there is a large and growing demand for information between years when data are actually collected. This demand is being met by government agencies and commercial data vendors, with private data vendors generally providing more detail and data for smaller geographic units than government agencies.

Demographers have long used population estimates and projections in the absence of actual data, and a variety of techniques are utilized to generate estimates and projections. Population estimates for states, MSAs, and counties are prepared each year as a joint effort of the Census Bureau and the state agency designated by each state governor under the Federal-State Program for Local Population Estimates (FSCPE). The purpose of the program is to standardize data and procedures so that the best quality estimates can be derived. Most states also generate population estimates and projections that are available through state agencies. However, these figures are often produced at irregular intervals, and thus may be quite dated. The reader is also encouraged to evaluate the quality of these data to the best of his or her ability. For additional information on population projections and estimates see Smith, Tayman, & Swanson, (2002).

In situations where the required demographic data are not available it may be necessary to generate "synthetic data" by means of estimates and projections.

Population estimates and projections generated by government agencies have historically been the only ones available. Today, however, a number of data vendors provide these figures. These vendor-generated data are often made available down to small units of geography (e.g., the census tract) and in greater detail (e.g., sex and age breakdowns) than government-produced figures. They offer the ability to generate estimates and projections for "custom" geographies (e.g., for a market area) not available for government-generated statistics. The drawback, of course, is that some precision is lost as one develops calculations for lower levels of geography and for subsets of the population. However, the ease of accessibility and timeliness of these vendor-generated figures have made them a mainstay for those requiring demographic data for various levels of geography.

Issues have been raised concerning the quality of the synthetic data produced by both government agencies and commercial data vendors. Data users typically need the latest information possible, and in an effort to be expedient the question of quality sometimes has become a secondary concern. Any evaluation of the quality of synthetic data requires an understanding of the currency and quality of the historical data being used as a basis for the estimates and projections. Furthermore, attention must be paid to the methods and assumptions utilized to generate the figures. If, for example, one assumes that population growth in an area is gradual and can be described by a simple mathematical function, population estimates and projections will be reasonably accurate as long as the assumptions hold. However, to the extent that an assumption is wrong, the (incorrect) mathematical function will yield inaccurate estimates and projections. While it is not possible to be aware of all the nuances of data quality and method, users are urged to evaluate underlying assumptions critically and to ascertain the accuracy of the synthetic data that are available. (Additional information on estimates and projections is provided in Chap. 9).

3.10 Sources of Data for Demographers

There are numerous sources of demographic data available today and the number of sources continues to grow. The sections below group these sources into four main categories: government agencies, professional associations, private organizations, and commercial data vendors.

It should be noted that the "products" available from these sources fall into two categories: (1) reports that summarize the data and (2) the actual data sets themselves. Historically, data access was essentially limited to summary tables provided by the organization, agency or vendor. Today, however, there is a trend toward providing the entire data set for use by planners and other data users. In reviewing the sources that follow, this distinction in format should be kept in mind. Exhibit 3.4 specifies sources of specific categories of data.

3.10.1 Government Agencies

Governments at all levels are involved in the generation, compilation, manipulation and/or dissemination of demographic data. The federal government, through the decennial census and related activities, is the world's largest processor of demographic data. Other federal agencies are major managers of data for the related topics of fertility, morbidity, mortality, employment and occupations, and migration. Various federal agencies compile and/or generate data of interest to demographers and often facilitate the dissemination of such data.

State and local governments are also major sources of data useful to demographers. State governments generate a certain amount of demographic data, with each state having a state data center for demographic projections. Vital statistics data can often be obtained in the most timely fashion at the state level, in fact. States vary, however, in the types and quality of data they generate. University data centers may also be involved in the processing of demographic data. Local governments may generate demographic data for use in various planning functions. City and county governments may produce population projections, while county health departments are responsible for the collection and dissemination of vital statistics data.

3.10.2 Professional Associations

In recent years, many professional associations have made an increasing amount of information on their members available to the research and business communities. Not only do such organizations have an interest in exchanging information with related groups, but they also have recognized the revenue generation potential of such databases. Some of the databases provide by professional associations include only basic information, while others offer a wealth of detail. In addition, some professional associations may conduct surveys or otherwise compile industry data that may be useful to demographers.

3.10.3 Private Organizations

Many private organizations (mostly not-for-profit) collect and/or disseminate demographic data. Voluntary associations often compile, repackage and/or disseminate such data. The American Cancer Society, for example, distributes morbidity and mortality data as it relates to its areas of interest. Some organizations, like Planned Parenthood, may commission special studies on fertility or related issues and subsequently publish this information.

Many organizations repackage data collected elsewhere (e.g., from the Census Bureau or the National Center for Health Statistics) and present it within a

specialized context. The Population Reference Bureau, a private not-for-profit research institute, distributes population statistics in various forms, for example. Some, like the American Association of Retired Persons (AARP), not only compile and disseminate secondary data but are actively involved in primary data collection, as well as the sponsorship of numerous studies that include some form of data collection.

3.10.4 Commercial Data Vendors

Commercial data vendors represent a fourth category of sources of demographic databases. These organizations have emerged to fill perceived gaps in the availability of various categories of data. These include commercial data vendors that establish and maintain their own proprietary databases, as well as those that reprocess and/or repackage existing data. For example, there are vendors that maintain databases that make this information available in a variety of forms. Major data vendors that do not necessarily create databases may incorporate demographic data into their business database systems. Some data vendors conduct major nationwide health consumer surveys. Some researchers have raised questions with regard to the quality of data produced by commercial vendors and with the lack of oversight related to confidentiality (Swanson, 2013).

3.11 Future Prospects for Demographic Data

The success of the demographic enterprise depends on the availability of accurate, timely and detailed data. Fortunately, the sources of demographic data have become more plentiful and more accessible over time. Various federal agencies post data in various forms on the Internet and make information available in a variety of formats (usually electronic today with few print reports now being generated).

The use of the American Community Survey for the collection of data originally collected through the decennial census relies on a smaller sample of the population than the one-in-six-household long form from the census, but the more frequent data collection improves the timeliness if not the accuracy of the data. It is anticipated that the Census Bureau will take advantage of non-census sources of data in the future, accessing data from other federal agencies (e.g., Social Security, Medicare files) and interfacing with non-government databases. It is also anticipated that use of sophisticated modeling techniques will become more common in an effort to close associated with gaps in traditional data collection techniques.

The acquisition of accurate, timely and detailed demographic data will continue to be a challenge. Various federal initiatives encourage more data sharing, and over time better access to data is anticipated. Here, too, data modeling will be increasingly important since the most common types of data related to many areas of

interest to demographers are never going to be compiled in any but very incomplete data sets. Persistent gaps in key data elements will require greater emphasis on modeling techniques for the generation of demographic data. Geographic information systems are expected to find an increasing range of applications in demography. Exhibit 3.4 summarizes the various sources for different categories of demographic data.

Exhibit 3.4: Selected Sources of Demographic Data

Information category	Source
Population data	
Size	ACS, Census, CPS, Vendors
Characteristics	ACS, Census, CPS, Vendors
Estimates and projections	Census, CPS, Vendors
Vital statistics	
Births	NCHS
Deaths	NCHS
Marriages	NCHS
Divorces	NCHS
Legal induced abortion	NCHS
Fertility	NCHS
Mortality	NCHS
Migration data	
Internal migration	ACS, Census, CPS, IRS
Immigration	ICE
Morbidity data	
Disease surveillance	CDC
Incidence/prevalence	NCHS
Health status	NCHS
Health risks	NCHS, BRFSS

Legend:
ACS = American Community Survey
BRFSS = Behavioral Risk Factor Surveillance System
Census = Decennial census
CDC = Centers for Disease Control and Prevention
CPS = Current Population Survey
ICE = Immigration and Customs Enforcement
IRS = Internal Revenue Service
NCHS = National Center for Health Statistics

Exercise 3.1: Data Available from the Census Bureau
The U.S. Census Bureau is the largest data collection organization in the world. Much of the data collected by this federal agency is useful for applied demography. A large portion of the data is collected through the decennial census and through the on-going American Community Survey. However, there are a number of other surveys administered by the Census Bureau of relevance to demographers. There are also registration systems and other means of tracking data utilized by the Bureau.

For this exercise, students should access www.census.gov and spend some time perusing the site and familiarizing themselves with the types of data that are available. Then, each student should compile a list of the various data elements that are available to allow them to develop an in-depth profile of their respective communities. Much of this will be the types of demographic data discussed this chapter. However, it is impossible to describe all of the data available, so students should develop a list of all of the types of data that might be useful in proofing a community.

References

Source: Gobalet, J., and S. Lapkoff, Lapkoff & Gobalet Demographic Research, Inc., www.demographers.com. Unpublished case study.

Smith, S. K., Tayman, J., & Swanson, D. A. (2002). State and local population projections: Methodology and analysis. *European Journal of Population, 18*(3), 303–305.

Swanson, D. (2013). Consumer demographics: Welcome to the dark side of statistics. *Radical Statistics, 108,* 38–46.

Additional Resources

Bureau of Labor Statistics (Department of Labor) website: www.bls.gov.

Census Bureau (Department of Commerce) website: www.census.gov.

Centers for Disease Control and Prevention (Department of Health and Human Services) website: www.cdc.gov.

ESRI website: www.esri.gov (for GIS applications to demography).

Immigration and Customs Enforcement (Department of Justice) website: www.ice.gov.

National Center for Health Statistics (Department of Health and Human Services) website: www.cdc.gov/nchs.

United States government statistical website: www.fedstats.gov.

Wombold, Lynn (2008). "Sample Size Matters: Caveats for Users of ACS Tabulations," *ArcUser* (Winter).