Chapter Two

Population Data

What Is a Population?
Types of Data
Data Sources
Data Quality
Conclusion
Focus: Census Data and the American Community Survey (ACS)
Methods, Measures, and Tools: Working with Data

DATA IS THE CORNERSTONE of demographic and population analyses. The existence of high-quality, publicly released data files enables much of this research, with the use of such data often accompanied by theoretical approaches commonly grounded in positivistic science, the goal of which is to verify (or falsify) empirical observations and to construct laws that can be generalized to a wide variety of models and theories. However, their use can also be problematic. In part, such data files are often considered incomplete. For example, they often miss details and the motivations for migration, immigration, and assimilation and instead rely upon empirically quantifiable notions of movement, acculturation, and statistical inference. Even the immigrant population is typically broadly defined and fails to distinguish between legal immigrants, illegal entrants, and refugees. Likewise, few data files detail the motivations for fertility choice. Perhaps not surprisingly, questions have arisen over the continued use of public data files and positivistic methods as a primary insight into population questions at the personal and societal levels.

Data differs in its content (what variables or constructs are included in the data), quality (how representative is it of the population), timeliness (what time period does it cover, or how related is it to specific events), coverage (geographic area) and availability (can the analyst access the data?). Given that each of these are important issues that can affect the analysis and interpretation of the data, it is useful to spend some time discussing alternate data sources. This

chapter presents and discusses different types of data. It begins by differentiating between populations and samples before discussing both qualitative and quantitative data, their sources, issues of data quality, and the benefits and costs of each type of data. The "Focus" section examines the US census and the American Community Survey (ACS), and the "Methods, Measures, and Tools" section discusses working with data.

WHAT IS A POPULATION?

Before going too far, we should define what we mean by population, as it is a word that can be used to describe various concepts, with a biologist's definition differing from that of a population geographer, by whom it is usually used to define a group of people. So far in this text, the concept has been generally used to describe the population of the world, a country, a city, or some other geographic unit. A population could equally represent people in a class at school, the population of juniors on campus, or the entire campus population. Regardless, each population has some boundary that defines who is included in the population (and equally important, who is excluded from it) and/or a common, shared characteristic (i.e., students in the class), so that the definition may be as precise as possible by either including or excluding individuals from the population. If, for example, our population is that of New York, we also need to specify what we mean by "New York." Without a geographical reference, answers that include the state, New York City, or the New York metropolitan area would be equally correct, yet each gives a very different answer. We must also consider the time period we are looking at. Are we, for instance, interested in the population of New York in 1900 or 2000, or somewhere between these two dates? The inclusion of time in the definition can make population a dynamic or changing concept.

While the intent is to carefully define what we mean by population, it is often difficult or impractical to work with a complete population, particularly if we are dealing with something as large as the population of a country. The numbers may be too large, the logistics too great, or the price tag too outrageous to count everyone on our own. Imagine, for instance, trying to count everyone in New York or another metropolitan area, and to also get them to answer questions on age, marital status, family size, education, mobility, and so on at the same time! The US Census Bureau does this every ten years with its decennial census (see "Focus"), but it is a huge, massive, and costly undertaking.[1] The 2010 census, for example, is estimated to cost over $14 billion, or about $16 per person, making it the most expensive census ever![2] As an alternative, population geographers will frequently use *samples* to represent the population.

Samples may be representative of the population, such as the American Community Survey or the Public Use Microdata Sample (PUMS), which is a 1 or 5 percent sample of the population based on the census. Samples such as these accurately reflect the structure and composition (age, gender, income, education, etc.) of a population and can be inflated with the use of sample weights to yield the actual population size for a specific geography. Samples may also be non-representative or purposive in nature, such that the researcher includes individuals of particular interest, such as new immigrants, older migrants, or women from a particular ethnic group. In such cases, results cannot be generalized or transferred to a larger population, as they are often specific to the group that is studied,[3] but fill the particular role needed by the researcher.

TYPES OF DATA

Generally speaking, there are two broad classes of data. Primary data refers to data that is collected by the researcher. It is usually collected one time only, is likely confined to a particular geographical area, and is typically a relatively small sample reflecting a particular problem or issue. While primary data can be costly and time-consuming for the researcher to collect and produce, it is usually flexible in that the researcher can define the questions and content of the survey along with the sampling frame—or how individuals are selected—to suit the particular needs or research questions.

Secondary data reflect data that have been collected by an organization, government body, or someone else using predefined questions, sampling frame, and geographic area. This data has also been typically checked, verified, and "cleaned," so that it is ready for public use. Advantages of such data sets are their (often) national representation and the detailed, robust sampling methodology that is used to construct the sample, so that data users can be assured of the representativeness of the sample: that it accurately represents the population it is based on. Sources of secondary data include, but are not limited to, formal statistical agencies such as the US Census Bureau, Bureau of Labor Statistics, or other national or international statistical agencies, with statistical agencies such as Statistics Canada or the US Census Bureau offering a number of different data sources, including censuses, labor force surveys, and health surveys. In the United States, data files include the Census and Current Population Survey (CPS) and longitudinal files such as the Population Survey of Income Dynamics (PSID), the National Longitudinal Survey of Youth (NLSY), and the ACS, which has replaced the decennial census "long form."

Both primary and secondary data sources can include qualitative and quantitative data. Qualitative data consists of nonnumerical information and may be

obtained through case studies, open-ended interviews, focus groups, participant observation, or diary methods. Participants may, for example, be asked for an oral history of their moves, including the reasons for moving and their destination choice and other related questions. Typically, these oral histories provide a rich understanding of the process in question but are also limited in terms of their ability to generalize findings beyond the sample or context of the analysis, given that they are often based on a small sample size. In contrast, quantitative data is numerical and includes counts, such as the number of people by age and gender in a specific area, measures of their educational attainment, place of residence and mobility data, and other socioeconomic or sociodemographic details. From these, rates, proportions, and other measures can be generated through statistical means to describe the population of interest.

DATA SOURCES

Geographers are often interested in such things as population structure and composition, transportation, population-environment issues, and population health. To understand, comment on, and offer solutions to these problems means that appropriate data is important. Where can population geographers turn to find the data and how "good" must data be to answer these questions? We can consider five main data sources: censuses, representative sample surveys, vital/civil registrations, indirect sources, and primary data that are collected by the analysts themselves.[4]

Census Data

The census—defined as the collection of demographic, economic, and social data pertaining to a particular time and country—is perhaps one of the best known and most used sources of population data. Counting or enumerating every individual in a population, the census offers a "snapshot" of a population at a particular time. In counting people, most censuses allocate them to their usual place of residence. These so-called de jure censuses differ from de facto censuses, which allocate people to their location at the time of enumeration. That is, if a person who works in Chicago, Illinois, but lives in Gary, Indiana were enumerated at work, they would be allocated to Chicago based on the de facto method but Gary based on the de jure method. De jure censuses are preferred as they provide a better indication of the permanent population in an area. In most cases, basic demographic and social characteristics of each person are also collected, including age, gender, marital status, household structure, educational attainment, and income. In addition, other household

characteristics may be collected, such as type of dwelling, occupation, and ethnic origin of respondents. In most cases, people are counted at their usual place of residence.

The widespread use of census data and other public data is due in large part to their validity and the degree of geographic, social, and economic detail embedded in the files (figure 2.1). Moreover, the growth of data has corresponded to increasing computational abilities and a refinement and broadening of the analytical tools used within population research, including the ability to test hypotheses through inferential techniques and gain insight into the causes and consequences of population movement. Not surprisingly, therefore, censuses represent a primary data source used by many population geographers. In the United States, the census has been carried out every ten years since 1790 (in years ending in 0), while Canada collects its census every five years (in years ending in 1 and 6). Both originated with the simple need to have a count of the population, but evolved to collect information relating to a variety of population characteristics. Most other nations also carry out censuses, although data quality and timing will vary.[5]

Representative Sample Surveys

Representative sample surveys are another source of population data, including national, regional, or state/province representative sample surveys that collect population information on individuals and/or households. A representative data source allows the user to draw generalized conclusions. These surveys do not have to focus exclusively on population topics to provide useful information. For example, in addition to running the census program, Statistics Canada runs a number of nationally representative data collection tools, including health, immigration, and youth surveys. While not meant to be population counts, these data sources provide background population characteristics, including age, location, gender, income, educational attainment, and household structure, to name a few. Other representative data sources that are frequently referenced by population geographers include other data files from the US Census Bureau, such as the ACS, which has been designed to replace the census long form (see "Focus") and the CPS. The CPS is a monthly survey of the American population and is the primary source of information on the labor force characteristics of the US population.

Vital Registrations

Vital registrations or civil registration systems record demographic events such as births, deaths (including cause of death), marriages, divorces, and population movements and provide yet another source of demographic data. Mortality

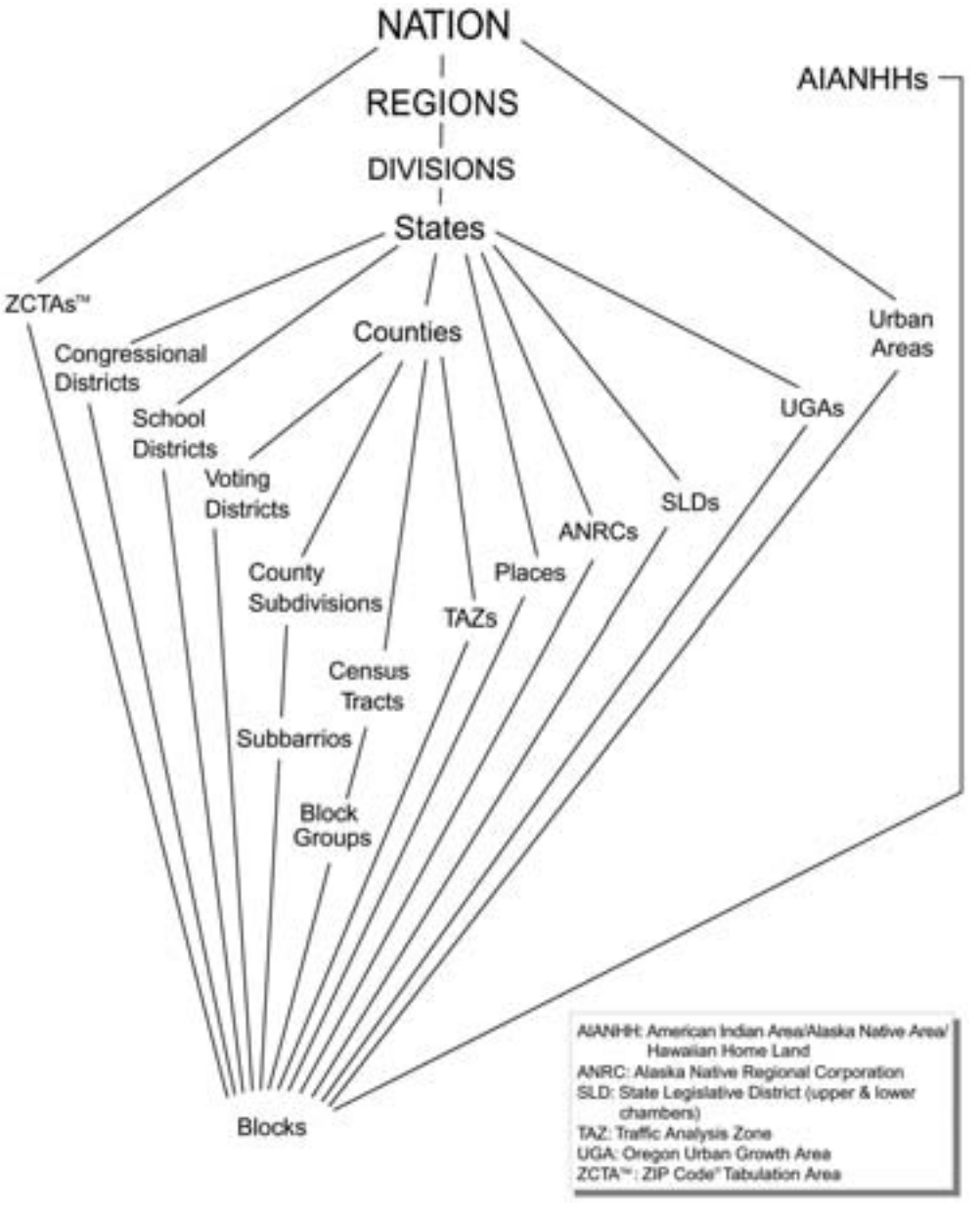

Figure 2.1 The Hierarchy of Census Geographic Areas.

Source: "Appendix A: Census 2000 Geographic Terms and Concepts," US Census Bureau, www.census.gov/geo/www/tiger/glossry2.pdf.

statistics are, for example, used in population projections to calculate the probability of surviving into a future period of time, while information on cause of death can be used to protect the health of communities. Most countries have legal provisions within their constitutions to ensure that vital events are recorded, although the type of information that is registered will vary from country to country, with vital registration systems more expansive (i.e., capturing population mobility in addition to births and deaths) in several European countries.

Other Secondary Data Sources

Beyond the census and its related products, numerous other secondary data sources are available for use by population geographers. In the United States, for example, agencies such as departments of Health and Education and the Bureau of Labor Statistics (BLS) also commonly collect statistics that either directly or indirectly provide population data. The Internal Revenue Service (IRS) also publishes migration data based on the addresses of tax filers, which allows the mobility of the tax-paying public to be tracked from year to year.[6] Immigration statistics (including refugee and asylee numbers) can be sourced from the Department of Homeland Security, and comparative international data can be sourced from Integrated Public Use Microdata Series (IPUMS) International, the PRB, and various United Nations (UN) agencies.[7] In addition, other agencies or organizations, such as the World Health Organization (WHO), the UN, and country-specific statistical agencies, collect and disseminate population data, and the Center for International Earth Science Information Network (CIESIN) has interesting data applications, including a "census by satellite."[8]

Analysts may also turn to less conventional secondary data sources for population information. One such use of these sources is illustrated by Foulkes and Newbold (2008),[9] who turned to data from local school boards and utility companies to measure mobility in small rural communities. In this case, data from the US Census Bureau was dated or was not available at the scale of analysis (small rural villages) used in the study. School board data was drawn from the Illinois State Board of Education's School Report Card file, which provided mobility and poverty data for each school district and individual school in the state, from which mobility rates could be calculated based on movement of students in and out of the school district. In addition, and as a potentially more inclusive source capturing mobility across all households (as opposed to just those with children in the school system), sewage-billing records were used to provide additional insight into local mobility, with change in billing name associated with movement into or out of the community. Although the use of this data allowed an analysis of population mobility among a subgroup of people

(rural poor migrants) that was otherwise missing from both the literature and from other data sources, it also illustrated the issues of using indirect sources, including data quality, comparability, replication, costs, and moral and ethical concerns.

Individualized Data Sets

In some cases, data from secondary sources is insufficient. Data may, for example, be outdated (as it was in the example above). Data may also miss, or lack sufficient numbers, for a particular population group or represent the wrong geographical scale. In each case, the researcher may be forced to construct her or his own data set(s). These "personalized" data sets offer a number of advantages, including enabling the researcher to select the sampling scheme, define the geographic scope and range of questions that are to be used, and include both qualitative and quantitative components in the research. Of course, there are also drawbacks to individual data sets. Most research questions or scripts will need to be vetted by institutional review boards, and researchers need to be aware of confidentiality and privacy issues. Although this is far from insurmountable, researchers need to ensure that quantitatively based samples will be sufficiently large or generalizable if they want to do statistical analysis or generalize to a larger population. Although the intent of qualitative studies is not to arrive at generalized conclusions, the collection, transcription, and coding of both quantitative and qualitative data can also be costly and time-consuming. The rewards, in terms of a data set that fits exactly the researcher's needs, can be large.

DATA QUALITY

Not all data sources are created equal, and they will differ in terms of their universality, quality, spatial scope, generalizability, validity, reliability, and replicability. In any data set, errors can be introduced in multiple ways, including the data collection process itself. For a census to be universal, everyone must be counted, but problems arise when some individuals or groups, such as the homeless, are difficult to count or refuse to be counted. For example, while there is always some undercount in any census, a post-census survey of the 1990 US census found that approximately four million people were missed. Populations were undercounted at different rates, with greater undercounting among the homeless, minority males in poverty, and Native Americans.[10] Underenumeration was particularly significant for cities. Apportionment of congressional seats and legislative redistricting was also affected by the under-

count. Local governments demanded (and received) population recounts, as federal transfers were reduced given the underestimated population sizes.

Respondents may also introduce errors into the data, affecting its quality. In some cases, respondents may not answer a question or set of questions, with questions regarding income often poorly answered. In other cases, individuals may attempt to deceive or provide answers that they feel are socially appropriate rather than their own. Famous among respondent (mis)information is age, with many tending to provide a younger age than reality. Likewise, questions related to past events are subject to "recall bias," such that facts, dates, or events are not recalled with complete accuracy and are dependent upon memory. In fact, most any question may be subject to some respondent bias, and a large literature exists on how best to construct and implement a survey.[11] Common issues also include the incorrect recording or transcribing of information or incorrect phrasing of questions. There was much discussion, for example, over the phrasing of the "come to stay" question for immigrants in the 1990 US census, with the question variously interpreted to mean the person's first entry into the United States, when they received permanent residency, or ultimately when they received US citizenship.[12] For researchers interested in immigration and adjustment, the timing differences can have significant implications! Finally, statistical agencies themselves may alter data quality by suppressing data, particularly for small populations or small areas, where data may be suppressed to protect confidentiality. The ACS (see "Focus" section), for example, will only release data for small geographic areas based on five-year rolling averages. In comparison, data for larger units will be released yearly.

For a geographer, space and its definition are often key. For those wishing to compare phenomena across space, representative data sources or census files offer a pragmatic solution: to generate space through a custom survey is probably time- and cost-prohibitive! But, if the analyst is interested in a particular locale or space, especially those for which formal data is inadequate or not available, individual one-time surveys are often best. For instance, a researcher interested in neighborhoods may need to use census-defined census tracts to proxy the neighborhood.[13] But, given the spatial variability of census tracts, particularly in rural or less dense areas, and the variability with which people define their neighborhood, this definition may be wholly inadequate.

As such, and despite the availability of a broad variety of data sources, caveat emptor (buyer beware) applies. Analysts should be careful to note whether data sources are nationally (or regionally) representative. Similarly, do vital registration systems capture all the data? For instance, births and deaths (particularly infant deaths) may not be reported, and causes of death may be mislabeled, incorrect, or missing. In general, completeness of registration is fairly high in the developed world; some South American countries, including Argentina,

Chile, and Colombia; and some Asian countries, including China, Sri Lanka, South Korea, and Japan. However, vital registration systems in most sub-Saharan African countries do not adequately or completely collect information on vital events.

CONCLUSION

In recent years, the abundance of secondary data sources has enabled geographers and other social scientists to make important contributions to understanding the demographic trends that shape our societies. The widespread use and availability of census and other public data are due in large part to their validity and the degree of geographic, social, and economic detail embedded in the files. The growth of data has also corresponded to new computational abilities and a refinement and broadening of the analytical tools used within population research. Moreover, given the geographer's interest in space and spatial relationships, the use of such large data files is therefore somewhat pragmatic: generating "space" through other means, such as individual one-time surveys, is typically cost- (or time-) prohibitive, as the sample size either needs to be large to adequately represent a particular location and/or needs to be replicated across space to capture spatial differences. This has not, of course, stopped researchers from constructing their own data sets or relying on qualitative data to understand demographic processes. In fact, these data sources should be seen as complementary rather than competitive, allowing different approaches and insights into population processes.

FOCUS: CENSUS DATA AND THE ACS

THE CENSUS

Censuses serve as a tool to count the population and ascertain its basic makeup. Many nations conduct censuses (counts) of their populations as tools to allocate government seats, funding, or other resources. Mandated by the United States Constitution, the first US census was conducted in 1790, and has been collected every ten years since then. The information collected by the Census Bureau is then used to distribute congressional seats and federal funds and to make decisions at every level of government.[1]

While the US census originated as a simple population count, it evolved over the years to include a variety of related population questions beyond age, gender, and address. In addition, a proportion of the population in the United States was frequently asked to complete the so-called census long form. Answered by approximately one in six people, the long form

included detailed socioeconomic and sociodemographic questions, including schooling and education, income, housing type, citizenship, immigration status, ethnicity, and race of all respondents in the household. This data is typified by the Public Use Microdata Sample (PUMS), published by the US Census Bureau. PUMS files are large, representative samples (1 or 5 percent) of the US population based upon the decennial census.

The main advantage of working with a file like the PUMS is the detail embedded in the files and the ease of generating population statistics such as migration numbers, flows, and net migration rates at a variety of spatial scales. Another significant advantage is the size. The fact that the PUMS is representative of the entire US population enables generalizability, validity, reliability, and replicability of analyses. Analyses can, for example, be replicated across space and time, such that changes in the structure or makeup of a population can be observed and tracked over time. For geographers interested in population mobility and migration, the PUMS also included a question on place of residence five years ago. First appearing in the 1940 census, it provided a measure of the mobility of the US population by contrasting place of residence at the time of the census with place of residence five years earlier, allowing a window into the migration habits of the population. In addition, the census includes information on nativity (immigrant or nonimmigrant) and period of entry into the country, which enables analysis of the mobility and economic characteristics of the foreign-born population.

While data files such as the PUMS provide detailed snapshots of the total population, secondary data sources have their limits. Oftentimes, key assumptions regarding relationships, measures, or definitions within the data must be made in order for the analysis to proceed. Because of their nature, analyses of secondary data files are often constrained by the data contained within them. For example, although the US census asks about immigrant status, it does not include information on whether an immigrant is legal or illegal. Thus, secondary data files typically offer little flexibility in defining variables or constructs.[2] Second, much of the information collected by the decennial census is quickly outdated. In effect, the census provides a snapshot of the population at one point in time on census day, while missing changes to the population that happened in the intervening years. Third, although long-term census data is available, with census files dating to 1850 available from the Integrated Public Use Microdata Series USA (IPUMS-USA),[3] the data is not always comparable. If the researcher is interested in comparing populations across time, changes in variable definitions (i.e., the changing nomenclature for cities or occupation codes) or the introduction/rephrasing/removal of questions to surveys complicates analyses.

THE AMERICAN COMMUNITY SURVEY

Given budget issues and congressional concerns with the invasiveness of questions in the census long form, the 2000 census was the last time the long form would be used.[4] The 2010 census will return to the more simple idea of a population count, collecting just name, age, sex, date of birth, race, ethnicity, relationship, and housing tenure. More detailed demographic information will be based on the ACS, which is perhaps the best known and one of the largest examples of a representative data source.[5] Meant to replace the information typically collected on the census long form,

the ACS provides current and up-to-date population estimates, along with estimates of demographic, economic, social, and housing characteristics of the US population at various levels of geography.[6] While it is not designed to count the population, it does provide an estimate of what the population looks like each year through the use of statistical sampling, surveying approximately one in forty households each year, with addresses selected at random and representing other addresses in the community (figure 2F.1).

For large geographies such as states or large metropolitan areas, population estimates will be released on a yearly basis. Given the sample size involved, population estimates for smaller geographies will be based on rolling averages. For small areas with a population less than twenty thousand, population estimates will be based on five-year averages. Areas with a population of twenty thousand to sixty-five thousand will be based on three-year averages,[7] with these rolling averages updated yearly. Similarly, population characteristics may pose problems for the release of the data, and releases may be forced to the five-year cycle. For instance, in areas with small ethnic or racial groups, the numbers may be too small to release yearly.

The ACS provides a number of advantages over the census long form, with the most significant advantage being the timeliness of the data. For instance, ACS data will mean that population counts will be updated on a yearly basis for large metropolitan areas (up to five years for smaller metropolitan areas) rather than every ten years based on the census. In addition, migration will be measured through a question asking place of residence one year ago. This means that migration researchers will be able to accurately evaluate yearly migration data on an ongoing basis. In comparison, the census bases its migration question on place of residence five years prior to census day, meaning that the actual migration might have happened up to ten years before the data is released! Second, the ACS will allow more up-to-date information on the characteristics of the population relative to the demographic event. For instance, the census measured sociodemographic and socioeconomic information at census day, but might be five years beyond the migration event and therefore outdated. In contrast, migration events and demographic or economic characteristics will be more tightly matched in the ACS, so that the person moving for educational reasons will be more closely correlated with the migration event itself. Third, the ACS will eliminate the gap in migration data. That is, migration data based on census returns has only been available for the second half of the decade, so that migrations in the first five years of each decade are missed. With the ACS, migration will be tracked every year.

For the population geographer interested in migration, the ACS also poses significant new analytical questions and problems.[8] For instance, by comparing place of residence on census day and five years prior, the old long form provided a consistent definition and timeframe of migration. The ACS, however, compares place of residence on the day the form is completed relative to where the respondent lived a year earlier. In this way, the window for migration (one year) is significantly less than with the long form (five years), and the timing of the migration is variable from one respondent to the next. So, two respondents in the same community may complete the ACS at two very differnt times in the same calendar year, and the relative timing of their migrations could reflect vastly different economic opportunities. In addition, comparison of

migration totals from the two sources will be problematic, since the number of migrants recorded over a five-year interval is considerably less than five times the one-year number[9] and the two totals cannot be easily reconciled.[10]

METHODS, MEASURES, AND TOOLS: WORKING WITH DATA

Good research should always be based on properly devised research questions that address gaps in the literature and are guided by appropriate theoretical foundations. At the same time, theoretical perspectives, methods, and data are also key to good research. While good data can help inform results, it does not guarantee "good" results. Likewise, researchers must use the appropriate methods to uncover what the data illustrates. Again, however, the choice of the research method or tool can alter results, even to the point of biasing outcomes and conclusions! In short, while we can use the analogy of "garbage in equals garbage out" (by substituting data/methods for garbage), we could have good data (methods) but still get garbage results if we haven't used it properly or haven't used the appropriate method.

THEORETICAL PERSPECTIVES

Regardless of the type of question to be addressed, theory is critical to providing a context for interpreting results and defining methods. If we take the example of migration, each individual migrant has his or her own reasons for migration, ranging from poverty and employment opportunities to amenities and health, with explanations encapsulated in various migration theories.[1] For instance, the human capital theory describes migration as an individual choice, with migrants treated as rational actors and able to look at various options, including origin and destination, wage rates, job security, and so on, while also accounting for the costs of migration. Alternatively, the *structural* perspective defines movement based on the social, economic, and political structures that shape people's lives, so that migrations are often forced.[2]

COLLECTING AND OPERATIONALIZING THE DATA

Following the research statement, one of the first tasks is to collect the appropriate data, a task that can be as complex as the actual use of the data. If a researcher is interested in a fairly specific segment of the population, such as young adults who are just leaving college, data collection may simply involve turning to existing data sources such as the census, downloading the data, and then defining the appropriate sample (i.e., by age). While the census is an easy data source to turn to, gathering data for the census is a complex task in itself. For instance, the 2000 census provided employment for some 860,000 temporary workers, and was billed as "the largest peacetime mobilization of resources and personnel."[3] Preparations for the 2010 census began almost immediately after the 2000 census was completed. As early as 2003, the Census Bureau was field

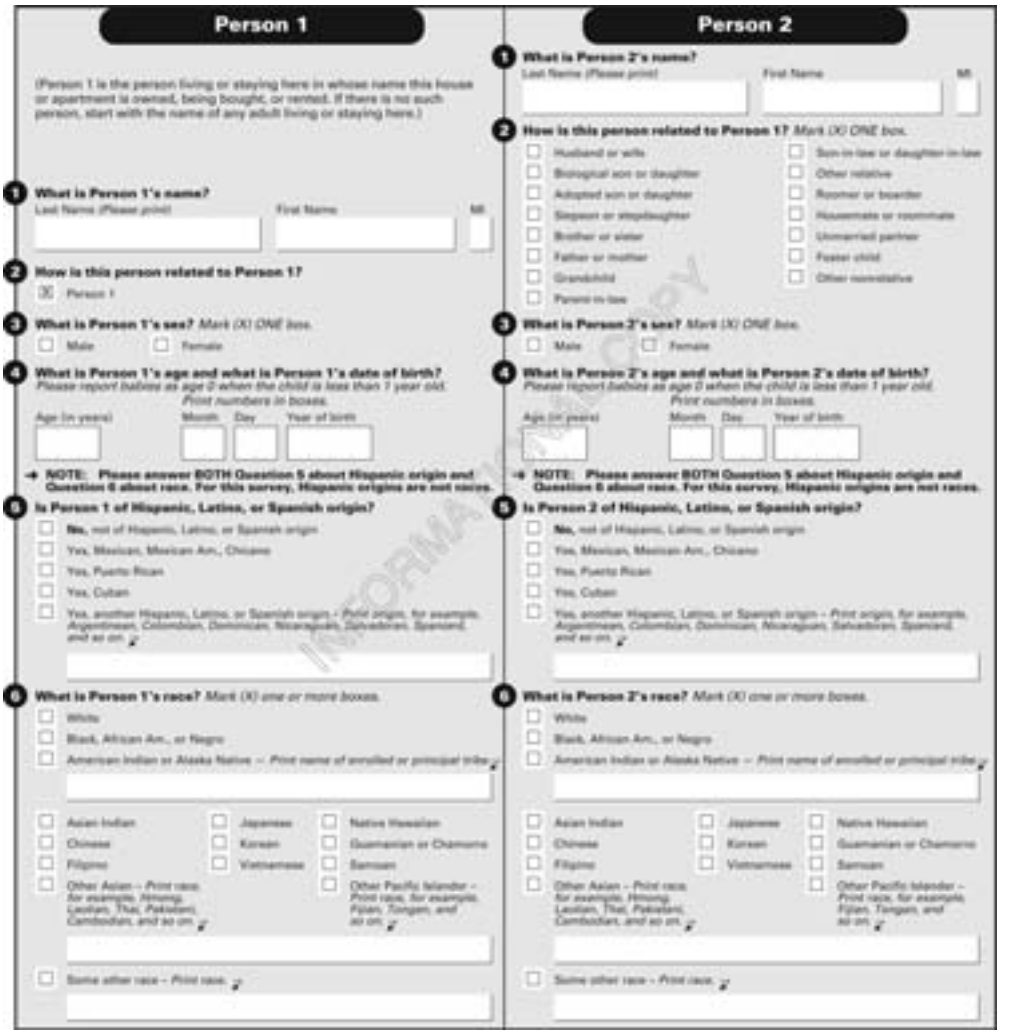

Figure 2F.1a The 2008 ACS Questionnaire.

Source: US Census Bureau.

Housing

Please answer the following questions about the house, apartment, or mobile home at the address on the mailing label.

1 Which best describes this building? *Include all apartments, flats, etc., even if vacant.*

- ☐ A mobile home
- ☐ A one-family house detached from any other house
- ☐ A one-family house attached to one or more houses
- ☐ A building with 2 apartments
- ☐ A building with 3 or 4 apartments
- ☐ A building with 5 to 9 apartments
- ☐ A building with 10 to 19 apartments
- ☐ A building with 20 to 49 apartments
- ☐ A building with 50 or more apartments
- ☐ Boat, RV, van, etc.

2 About when was this building first built?

- ☐ 2000 or later – *Specify year:*
- ☐ 1990 to 1999
- ☐ 1980 to 1989
- ☐ 1970 to 1979
- ☐ 1960 to 1969
- ☐ 1950 to 1959
- ☐ 1940 to 1949
- ☐ 1939 or earlier

3 When did PERSON 1 (listed on page 2) move into this house, apartment, or mobile home?

Month Year

A *Answer questions 4 – 8 if this is a HOUSE OR A MOBILE HOME; otherwise, SKIP to question 7a.*

4 How many acres is this house or mobile home on?

- ☐ Less than 1 acre → *SKIP to question 6*
- ☐ 1 to 9.9 acres
- ☐ 10 or more acres

5 IN THE PAST 12 MONTHS, what were the actual sales of all agricultural products from this property?

- ☐ None
- ☐ $1 to $999
- ☐ $1,000 to $2,499
- ☐ $2,500 to $4,999
- ☐ $5,000 to $9,999
- ☐ $10,000 or more

6 Is there a business (such as a store or barber shop) or a medical office on this property?

- ☐ Yes
- ☐ No

7 a. How many separate rooms are in this house, apartment, or mobile home? *Rooms must be separated by built-in archways or walls that extend out at least 6 inches and go from floor to ceiling.*

- *INCLUDE bedrooms, kitchens, etc.*
- *EXCLUDE bathrooms, porches, balconies, foyers, halls, or unfinished basements.*

Number of rooms

b. How many of these rooms are bedrooms? *Count as bedrooms those rooms you would list if this house, apartment, or mobile home were for sale or rent. If this is an efficiency/studio apartment, print "0".*

Number of bedrooms

8 Does this house, apartment, or mobile home have –

	Yes	No
a. hot and cold running water?	☐	☐
b. a flush toilet?	☐	☐
c. a bathtub or shower?	☐	☐
d. a sink with a faucet?	☐	☐
e. a stove or range?	☐	☐
f. a refrigerator?	☐	☐
g. telephone service from which you can both make and receive calls? *Include cell phones.*	☐	☐

9 How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of this household?

- ☐ None
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6 or more

10 Which FUEL is used MOST for heating this house, apartment, or mobile home?

- ☐ Gas: from underground pipes serving the neighborhood
- ☐ Gas: bottled, tank, or LP
- ☐ Electricity
- ☐ Fuel oil, kerosene, etc.
- ☐ Coal or coke
- ☐ Wood
- ☐ Solar energy
- ☐ Other fuel
- ☐ No fuel used

Figure 2F.1b The 2008 ACS Questionnaire.

Source: US Census Bureau.

testing questions via dress rehearsals, and by 2008 it had begun recruiting workers and updating address lists for the actual census.

For others, data must be collected through primary means such as surveys or interviews. If the latter, the sample must be identified and recruited (which individuals are asked to complete the survey and what is the sample structure—is it random? Snowball?, etc.)[4] and then asked to complete the interview, survey, or whatever data collection tool is being used. Typically, if the researcher wants results that are generalizable, then the sample must be random and representative of the population of interest. In other cases, the researcher may want to oversample particular communities or groups to ensure adequate information, while randomness may not be an issue in other cases. Collected data must then be entered or transcribed and checked for entry errors. If it is quantitative data, it should be also be checked for its representativeness of the population, which is usually done by comparing sample characteristics such as age and gender, along with such attributes as education and income, to known population values, such as those drawn from a census.

After this, we are nearly ready to start working with the data. So, how do we operationalize the data and what analytical methods are best? If we are examining population movement, how we define population movement is critical and depends on the research questions as well as the data. For instance, the international migration/immigration literature distinguishes *temporary immigrations*, such as short-term relocations; *transnationalism*; or *permanent immigration*. Likewise, the domestic migration literature also distinguishes between seasonal moves, such as the seasonal movement of "snow birds" between colder and warmer climates, and local moves (i.e., city), regional moves (i.e., county), or moves between states/provinces. Other issues, including the length of interval over which movement is captured (critical for looking at temporary migration); the size, shape, and characteristics of the receiving and sending regions; and the composition of the sample population, affect the analysis. As such, researchers must clearly define the population of interest.

METHODS

Research methods need to be defined, with researchers able to choose between a variety of different methods, selecting that which is best for their data. Qualitative data, for example, demand qualitative techniques, including the coding and interpretation of common themes or issues in the data,[5] with this analysis embedded within appropriate theoretical perspectives. For example, Strauss and Corbin suggest proceeding through open, axial, and selective coding.[6] Open and axial coding involve line-by-line coding of the data (microanalysis). During open coding, the data is examined and initial themes and concepts are generated. This process involves reading through each interview line by line looking for themes and concepts. Axial coding reexamines the themes and concepts identified within open coding by identifying their interrelationships (i.e., the networks and hierarchies that exist among and between them). Axial coding results in the development of an array of interrelationships between the various themes identified during open coding. In the final stage, selective coding is used to integrate and refine the categories and subcategories identified through open and axial coding. This requires the identification of central catego-

ries that represent the main theme of the research, which is defined as one that "has the ability to pull the other categories together to form an explanatory whole."[7] These central categories form the larger theoretical framework.

For the quantitative geographer, a series of tools are also available. Descriptive statistics, for example, including the calculation of means, standard deviations, or basic cross-tabulations, characterize the data and allow its exploration. Such descriptive analyses also provide a way to ensure that the sample is representative of the population of interest. While this stage is less crucial when using data files such as those from the US Census Bureau or Statistics Canada, which are representative of the population, it is key for researchers using data they have collected themselves. Following the initial description of the data, the analyst may turn to other methods and techniques, including inferential and multivariate statistics. Geographic information systems and spatial analytical techniques, including mapping of data, understanding geographic trends in the data, and searching for clustering or hot spots, are also widely used. All bring statistical significance to the analysis and allow a better understanding of the data. Many of these techniques are discussed elsewhere in this text.