U.S. Census data and ArcGIS: an orientation

The following is taken from Peters and MacDonald, *Unlocking the Census with GIS* (ESRI Press, 2004).

- *No matter what the data source, always check for possible errors (e.g., add up the "Race" and gender categories to make sure they sum to POP2000, and add up the housing tenure categories to make sure they sum to HSE_UNITS).*
- *And check errata (e.g., U.S. Census) when they are available.*

Basic organization of U.S. Census data

- *Census block*: lowest level of tabulated census geography. Defined by roads, natural features, or politico-administrative boundaries. Typically has a population of about 85 people. Contains relatively little census information–including no socioeconomic data—due partly to issues of confidentiality.
- *Block groups*: aggregations of census blocks that are the lowest level for which sample census data—including a bit of socioeconomic data—is available. Range from 600 to 3000 people.
- *Census tracts*: aggregations of block groups. Meant to be fairly homogenous and permanent areas with similar economic and demographic characteristics and living conditions. Range from 1500 to 8000 people. The smallest unit that has ample socioeconomic data (based on the census's "sample forms").
- *Counties or country equivalents (e.g., parishes, boroughs, "census areas," "incorporated places")*: aggregations of census tracts.
- *States*: aggregations of counties and county equivalents.

SF1-SF4 & PL

- SF1 and SF2 contain data on the questions asked of all households: SF1 has the most geographic data (down to the block) and SF2 has the most socioeconomic, housing data, and other data (down to the block group or tract).
- SF3 and SF4 have data from questions asked of a sample of households only. SF3 has the most geographic data (down to the block group or tract) and SF4 has the most other data.
- PL (Redistricting data, Public Law) contains information on race and whether of Hispanic origin, based on the Census 2000 short form. There are four detailed tables.

FIPS (Federal Information Processing Standards) numbering system

- *State*s*:* two-digit code numbers (e.g., 22=Louisiana).
- *Counties or county equivalents*: three-digit codes attached to state codes (e.g., 071=Orleans Parish).
    - FIPSSTCO combines the state and county codes.
- *Census tracts*: six-digit codes attached to state and county codes.

    - A tract that has been divided into multiple tracts over time will usually have numbers in the decimal place positions of the tract code (e.g., 320.01 and 320.02).
- *Block groups and blocks*: numbered consecutively from "1," attached to census tract codes.
- STFID combines the state, county, tract, and block codes.

- In general, most census data are aggregated to a specific geographic scale, called the "summary level," with each level having its own summary level code (e.g., 140=state-county-census tract, 150=state-county-census tract-block group; see Peters and MacDonald, page 31).

The extended census geographic hierarchy

- *Places*: cities, towns, villages, etc., have a special five-digit code based on the alphabetical order of the place's name within a state.
- *Metropolitan areas*: counties or sets of counties with a population over 100,000 (75,000 in New England) and a central city population of at least 50,000. Outlying counties are included if they have enough workers that commute to the central city for employment and if they have high enough population density, urban population, and population growth.
- *Urban/rural areas*:
    - *Urban areas*: densely settle areas consisting of blocks or block groups with at least 1000 people per square mile, surrounding blocks or block groups with 500 persons per square mile, less densely settle blocks or block groups that form connections across more densely population blocks or block groups.
    - *Rural areas*: people and territory located outside the above areas.
- *Purpose-defined districts*: school, state legislative, and voting districts.
- *Zip code tabulation areas(ZCTA)*
- *Traffic Analysis Zones (TAZ)*
- *Public Use Microdata Areas (PUMAs)*: one-percent sample of raw long-form but with confidentiality screening for nation, states, and some substate areas, and five-percent sample for state and substate areas. One percent Super Pubic Use Microdata Areas (Super-PUMAs) and 5 percent PUMAs.

Major census data tabulations

- *100% data*
    - *Redistricting Data Summary File*: population counts; blocks.
    - *Demographic Profile*: selected population and housing characteristics; tracts.
    - *Congressional District Demographic Profile*: as for Demographic Profile; congressional districts.
    - *Summary File 1 (SF1*): counts and cross tabulations on short-form questions; blocks.
    - *Summary File 2 (SF2)*: like SF1, but with detailed breakdowns by "race" and some ethnicity; tracts.
- *Sample data*
    - *Demographic profile*: selected population and housing characteristics; tracts.
    - *Congressional District Demographic Profile*: selected population and housing characteristics; tracts.

- *Summary File 3 (SF3)*: social, economic, and housing characteristics; block groups and tracts.
- *Summary File 4 (SF4)*: like SF3 but with detailed breakdowns by "race" and some ethnicity; tracts.
- *Public Use Microdata Samples (PUMS)*: see PUMAs above.

Choosing census data

- Choose a table making sure you understand the universe from which it is drawn and the available levels of aggregation.
- Find the constitutent data dictionary reference names of the variables.

Downloading census data and maps

- TIGER/Line files (Topologically Integrated Geographic Encoding and Referencing system).
  - Raw form: www.census.gov/geo/www/tiger/index.html
  - Download and unzip; perhaps use ArcToolBox Wizard to convert data to GIS formats
  - Note the unprojected coordinates of the files.
  - *Census Feature Class Codes (CFCCs)*: allow users to identify individual features.
    - *Class A*: roads.
    - *Class B*: railroads.
    - *Class C*: miscellaneous ground transportation.
    - *Class D*: landmarks.
    - *Class E*: physical features.
  - E.g., how to find the income and demographic characteristics of people living next to highways in a particular county: download the road and block groups data; select the roads that are highways (probably A11 to A18); and find the black groups that intersect with the road types.
- Another option: use the U.S. Census's *American FactFinder* web site.
  - One approach is to download data in Excel format, save a preliminary version with renamed field columns but conserved field descriptive labels, and then save a final version as a DBase4 file.
- Another option: download free, already processed census data.
  - www.geographynetwork.com/data/**tiger**2000/
  - www.**esri**.com/data/download/census2000_**tiger**line/
  - But TigerLine files are restricted to SF1 and PL.
  - Or data web sites maintained by states, cities, or universities.
- Another option: a commercial, customized census data service.

Displaying the data

- Use a *table join* based on the "key" or "key field": e.g., STFID.
  - Data/Export Data/save as a shapefile or geodatabase feature class.
- In more complex cases, you may have to use ArcGIS to create key fields (e.g., for data downloaded from *FactFinder*).
  - E.g., you may have to extract the STFID code from the GEO_ID field, which involves converting the text file to database format using Access and then adding the new field via ArcGIS (see Peters and MacDonald, pages 67-69).

Finding digital maps: see Peters and MacDonald, pages 74-84.

Principal variables for study of social inequality & housing (adapted from Peters and MacDonald, *Unlocking the Census with GIS*).

Always check data for possible errors (e.g., add "Race" categories to make sure that they sum to the total for POP2000).

District demographic variables
Median age; % highest level of education completed; % black, hispanic, asian, white population; % population that lived elsewhere 5 years ago; % non-U.S. citizens; % foreign-born population; % population that lived in another country 5 years ago; % population that lived in another state 5 years ago; % ages 16-19 enrolled in school; % age 16+ in labor force (and do by race-ethnicity and gender); % age 16+ employed; usual means of transportation to work; usual travel time to work; usual travel time to work by usual means of transportation; % household income per capita; % households in relation to poverty line. When possible, cross-tab most of these variables by race-ethnicity, gender, and by female-headed families/households with children vs. other families/households).

Household composition variables
Median household size; % households with subfamilies; % female-headed households with children; % households with members 5 years old or less; % households with members 65+ years old; householder's (i.e. Census-defined head's) age; householder's highest educational level completed; % households that regularly speak a non-English language at home; % of households where no one over 14 speaks English only and no one over 14 speaks English "very" well (i.e. "linguistically isolated" households). When possible, cross-tab most of these variables by race-ethnicity).

Housing variables
% ownership occupancy; % renter occupancy; % vacant (best to exclude seasonal vacancies); unit type; year built; year householder moved in; rent asked; gross rent; gross rent as % income; mortgage status of owned homes; monthly owner costs; owner costs as % income; # rooms; # bedrooms; # occupants per room; # occupants per bedroom; overcrowded unit (1.5+ persons per room; cross-tab by owner and renter occupancy and by race-ethnicity); telephone service in home; vehicles available at home; complete plumbing facilities; complete kitchen facilities. When possible, cross-tab most of these variables by race-ethnicity, gender, and by female-headed families/households with children vs. other families/households).