Massachusetts Institute of Technology
Department of Urban Studies and Planning

11.520: A Workshop on Geographic Information Systems

11.188: Urban Planning and Social Science Laboratory

Making Sense of the Census

October 12, 2005

Overview

- What is it and why do we care?
- How the data are collected?
- What data are available?
- Introduction to Census geography and summary levels
- A Quick Look at the Census documentation
- A Quick Look at some sample data

What Is It and Why Do We Care?

- Mandated by the Constitution of the United States
- The modern census of population and housing was established in 1940 with the incorporation of the housing component and the introduction of sampling techniques for the long form
- Conducted every ten years
- Attempts an actual count of population categorized by various criteria
- The only source for demographic data with a wide geographic scope
- The most reliable and detailed information for describing local areas: neighborhoods, cities, counties
- The most consistent source of time series demographic data available
- U.S. Congressional representatives are apportioned based on census counts. Federal dollars for schools, employment services, highway assistance, housing construction, hospital services, programs for the elderly, etc. are all distributed based on census figures.

How the Data Are Collected

- Collected from households through a mail survey conducted every decade
- For the **2000 Census** more than 285,000 census takers and support personnel accounted for the 118 million households and 275 million persons in the United States.
    - 2000 Census Home Page
- Two different census questionnaires are distributed:
    - short-form questionnaire contains questions asked of everyone (summarized in Summary Tape File 1 (STF 1) for 1980 and 1990, Summary File (SF 1) for 2000)
    - long-form questionnaire contains questions asked of a population sample (1/6 households)
      (summarized in Summary Tape File 3 (STF 3) for 1980 and 1990, Summary File 3 (SF 3) for 2000)
- The long form is being replaced in the 2010 Census by the American Community Survey. This program will survey homes every month and provide updated statistics every year instead of every ten years. The program begins in 2003.

What's Included: Information on Population, Employment and Housing Characteristics

- **Short Form: 100% Count (STF 1/SF 1)**

Population Characteristics	**Housing Characteristics**
Age	*Tenure*
Gender	*Value or Contract Rent*
Race	*Vacancy Status*
Hispanic Origin	*Number of Rooms*
Marital Status	*Units in Structure*
Household Type	*Congregate Housing*
Household Relationship	

- Sample Short Form from 2000 Census

- **Long Form: Sample Counts (STF 3/SF 3)**

Population Characteristics	**Housing Characteristics**
Social Characteristics	*Age of Housing*
Education	*Heating Fuel*
Citizenship	*Facilities*
Ancestry	*Vehicles*
Language	*Mortgage Status*
Disability	
Children	
Place of Birth	
Economic Characteristics	
Income	
Labor Force Status	
Employment	
Place of Work	
Public Assistance	
Retirement Income	

  - Sample Long Form from 2000 Census

- **Why We Need to Know the Two Components**

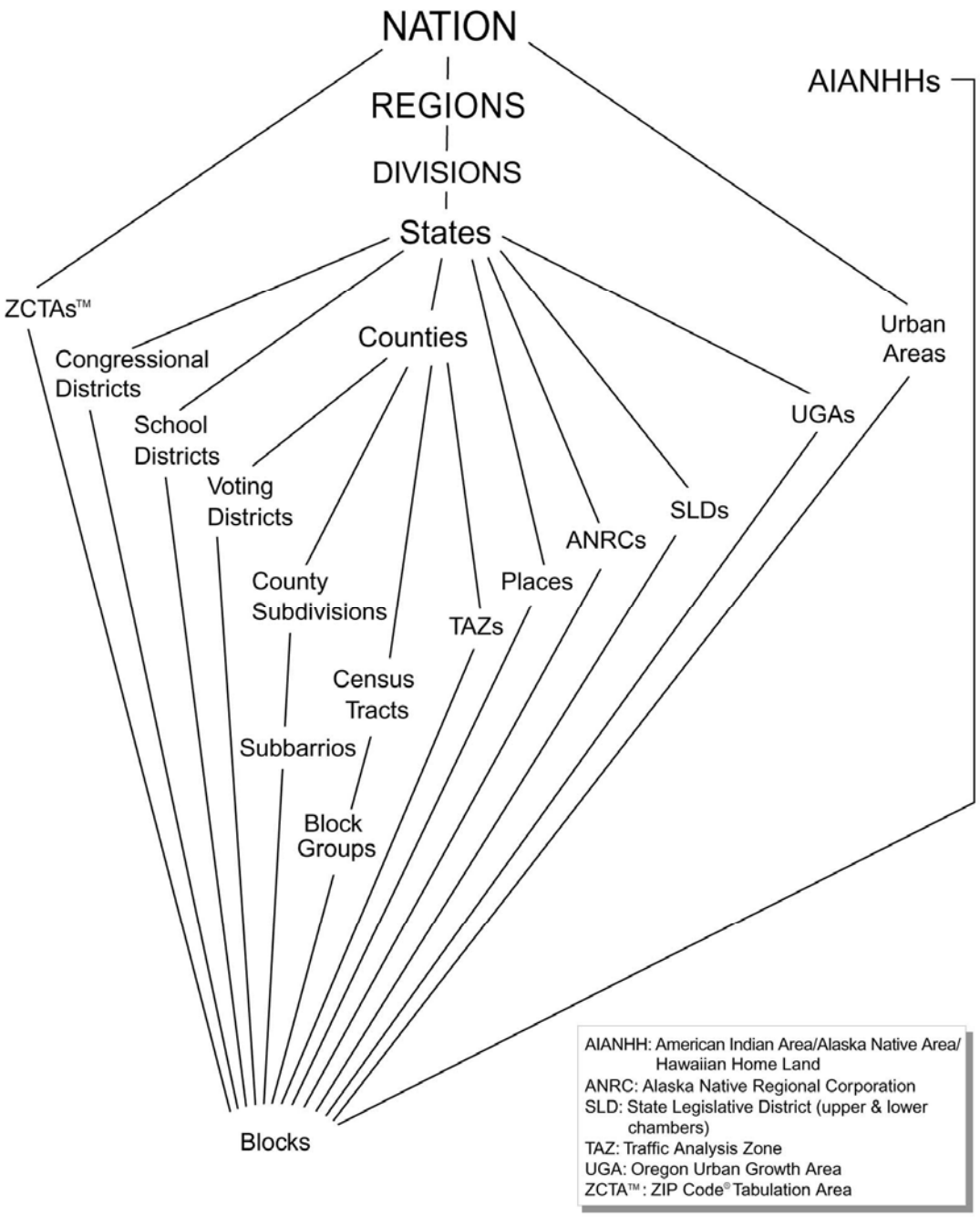
  - Accuracy of the data varies and counts differ (Why?)
  - It helps us to understand how the data are organized in Summary Tape Files (STFs)

Census Geography and Summary Levels

The Census organizes and aggregates data into a series of geographic hierarchies

- **Overview**

Standard Hierarchy of Census Geographic Entities (from *Census 2000 Summary File 1 Technical Documentation*, prepared by the U.S. Census Bureau, 2001, p. A-25)



Summary Level	Geographic Unit
010	United States
020	Region: Northeast (NE), Midwest (MW), South (S) and West (W) Regions
030	Division: Northeast: New England, Mid Atlantic Midwest: East North Central, West North Central South: South Atlantic, East South Central, West South Central West: Mountain, Pacific
040	State (includes Washington, D.C. & Puerto Rico)
050	County
060	County Subdivision
070	Place
080	Census Tract / Block Numbering Area (average 4,000 persons)
090	Block Group (average 1,000 persons)
100	Block (average 85 persons)

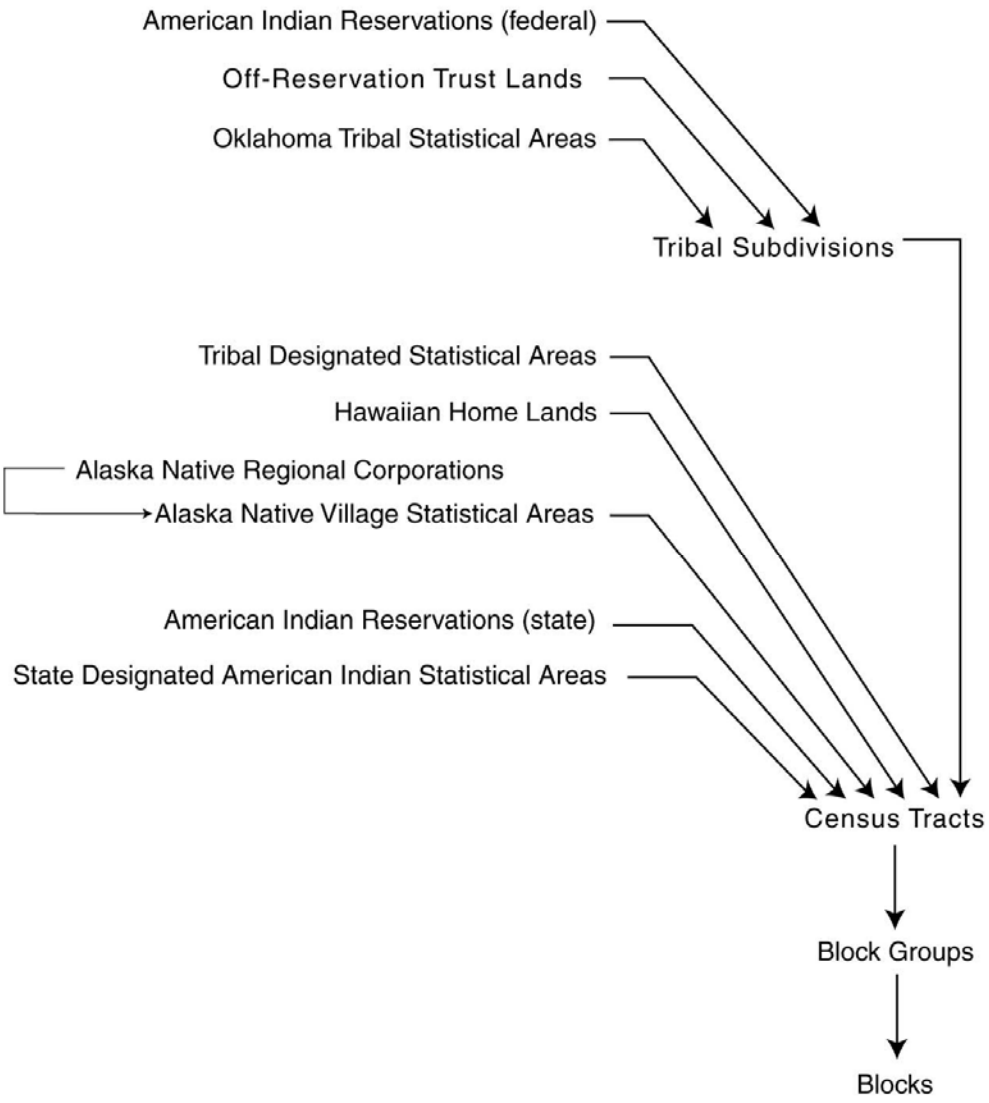
- **State-County-Tract-Block Group Nesting**

Summary Level	Geographic Unit
040	State (includes Washington, D.C. & Puerto Rico)
050	County
140	Census Tract
150	Block Group

- **Supplemental Geographic Areas**

Summary Level	Geographic Unit
400	Urbanized Areas
300	Metropolitan Areas (MSAs, CMSAs)
200	American Indian and Alaska Native areas
800	ZIP codes

Hierarchy of American Indian, Alaska Native, and Native Hawaiian Entities (from *Census 2000 Summary File 1 Technical Documentation*, prepared by the U.S. Census Bureau, 2001, p. A-26)

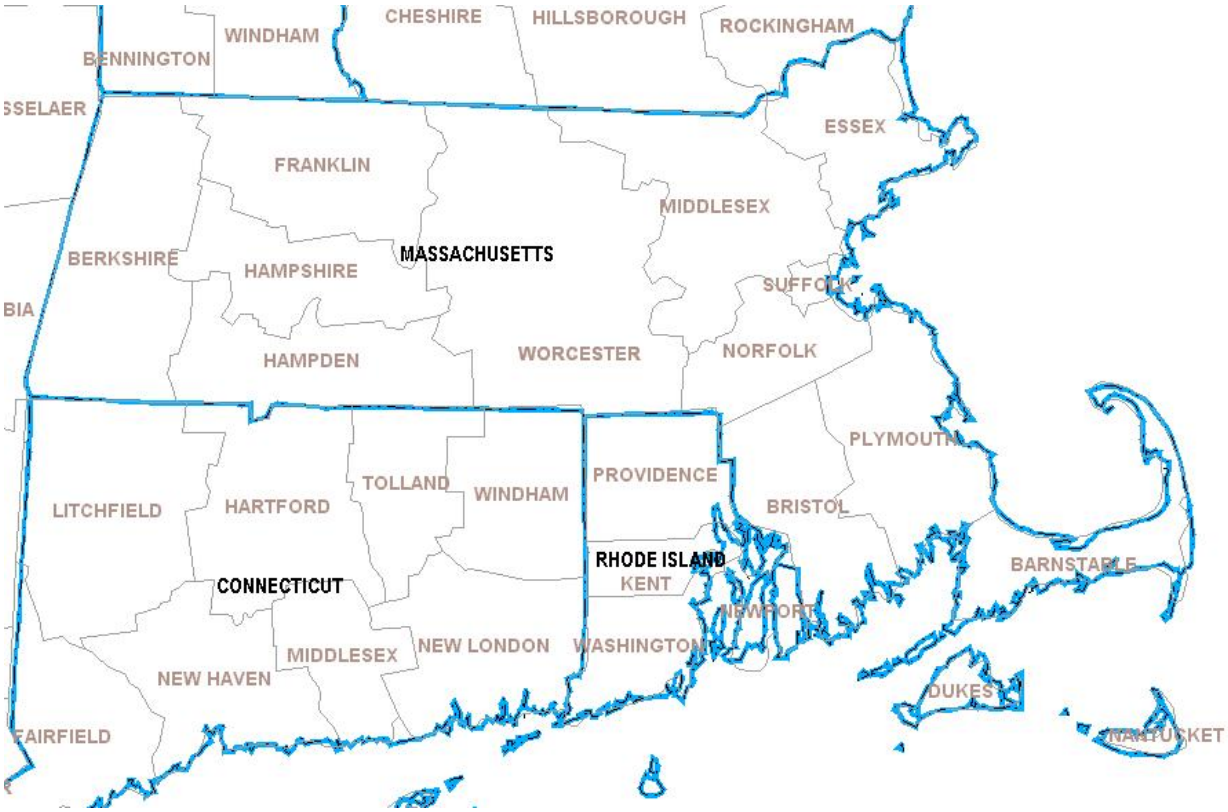


- **A Visual Look at Census Geography**
    - **Continental United States (Regions in blue; Divisions in green; States in brown)**

- **Counties**


WASHINGTON
OREGON
CALIFORNIA
IDAHO
NEVADA
MONTANA
WYOMING
UTAH
ARIZONA
COLORADO
NEW MEXICO
NORTH DAKOTA
SOUTH DAKOTA
NEBRASKA
KANSAS
OKLAHOMA
TEXAS
MINNESOTA
IOWA
MISSOURI
ARKANSAS
LOUISIANA
WISCONSIN
ILLINOIS
MICHIGAN
INDIANA
OHIO
KENTUCKY
TENNESSEE
MISSISSIPPI
ALABAMA
GEORGIA
FLORIDA
SOUTH CAROLINA
NORTH CAROLINA
VIRGINIA
WEST VIRGINIA
PENNSYLVANIA
NEW YORK
MARYLAND
NEW JERSEY
CONNECTICUT
MASSACHUSETTS
MAINE
Pacific
Mountain
W. N. Central
E. N. Central
W. S. Central
E. S. Central
Mid-Atlantic
New England
South Atlantic

- **A Closer Look at Southern New England Counties**



- **Tracts (black lines) and Block Groups (green lines) in Suffolk County, MA**

- **Census Geography Concepts**
  - The census block is the basic level
  - Confidentiality must be maintained, and data about individual persons and households are not revealed
  - More detailed data are provided for higher levels of geography (Why?)
  - Many, but not all, items are available at multiple summary levels

- **Potential Problems**
  - The same geographic name is used for summary levels corresponding to different aggregations
  - Geographic areas at lower levels may be subdivided by higher levels of geographic units
    - E.g., a census tract may be split by town boundaries
  - The same variable names are used for different variables in the STF/SF 1 and STF/SF 3
  - The way variable values are encoded makes identifying the meaning of variables difficult
  - ZIP codes do not overlay other units cleanly
  - Geographic boundaries change with time, making time-series analysis difficult.

- **Obtaining Census Geographic Boundary Files for Use in a GIS**

  ArcView shapefiles and ArcInfo coverage formats are readily available for 1990 and 2000 Census geography boundaries

  - Boundary files from the U.S. Census Bureau
  - Redistricting TIGER 2000 from ESRI's Geography Network

Census Summary Files

The most useful files distributed by the Census Bureau are the Summary Tape Files (now renamed simply Summary Files) that aggregate the individual census forms to various levels of census geography.

The Census Bureau distributed the 1990 Census files as DBF files on CD-ROMs. In what looks like a recent development, the Census Bureau has posted the contents of many 1990 CD-ROMs online. These are available via HTTP and FTP. An online forms-based interface called 1990 Census Lookup is available. Now, American FactFinder provides another forms-based online interface.

In fact, the **1980** STF 1 and STF 3 are now online! You can obtain the 1980 STF 1 via HTTP or FTP and the 1980 STF 3 via HTTP or FTP. Documentation is available from the Odum Institute for Research in Social Science.

The Census Bureau is distributing the 2000 Census files on CD-ROMs, DVD-ROMs in a proprietary format and online in flat ASCII format via HTTP and FTP. American FactFinder provides a forms-based online interface.

- **STF/SF 1: 100% count data from the short form**
  For the 2000 Census, the SF 1 files encompass all summary levels.
  For the 1990 Census, the STF 1 files came in four varieties:
  - A: States and subdivisions to the block group level
  - B: Block level
  - C: Entire U.S. and major subdivisions
  - D: Congressional Districts

- **STF/SF 3: Sample data from the long form**
  For the 2000 Census, the SF 3 files will encompass all summary levels.
  For the 1990 Census, the STF 3 files came in four varieties:
  - A: States and subdivisions to the block group level
  - B: 5-digit ZIP codes
  - C: Entire U.S. and major subdivisions
  - D: Congressional Districts

The 1980 STF 1 and STF 3 files had varieties similar to those of the 1990 Census.

A Quick Look at the Census Data and Documentation

1980 Census

- Overview from SUNY Albany's Center for Social and Demographic Analysis
- Data sets available from IPCSR (For MIT affiliates: Information on Accessing ICPSR Data)

1990

- STF 3A Variable Locator
- Table Definitions (Matrix)
- State/County FIPS Codes
- Census Data at the Center for Disease Control and Prevention

2000 Census

- American FactFinder
- Public Law 94-171 (PL 94-171)
  - Home Page
  - Documentation
  - Help on Using Browser Software on the CD-ROM
  - Data
- Summary File 1 (SF 1)
  - Home Page
  - Documentation
  - Help on Processing Data Files in ASCII Format
  - Data
- Summary File 2 (SF 2)
  - Documentation
  - Help on Processing Data Files in ASCII Format
  - Data
- Summary File 3 (SF 3)
  - Documentation
- Summary File 4 (SF 4)
  - Documentation

Censuses in Other Countries

- International Statistics Agencies

More Information About the 2000 Census

Data Release Dates
Subjects Areas of Questions Asked

Example: Let's find the unemployment rates for Cambridge area block groups

- **How can we measure unemployment rate**: how about the fraction of adults aged 16 or over who are in the labor force and are unemployed (during the sample week in April 1999)
- **Find the relevant SF3 census 2000 variables**: we use the SF3 technical documentation (Ch. 3) to find variable P43: employment status by sex, and the name of the text file that includes the raw data for this variable (ma00004.uf3)
- **Find and download the zipped datafile** that contains P43 for Massachusetts as an ASCII 'flat file' - this file is called: ma00004.uf3
- **Find and download the zipped datafile** that contains the geographic identifiers for Massachusetts - this file is called: mageo.uf3
- **Find and download the MS-Access templates** that will let you pull the ASCII plain-text data into MS-Access:
  - Explained in the 'readme.txt' file in the same directory as the zipped data files. Note, that readme.txt also includes the cross-referencing of the census variables (such as P43...) with the text file that bundles the data (such as ma00003.uf3).
  - The zipped template for MS-Access 2000 is here: http://www.census.gov/support/2000/SF3/Acc2000.zip
- **Import the relevant Mass data into Access tables**
  - rename the unzipped text files to end in 'txt'
  - Use the File/Get-external-data/Import option in MS-Access, with the file type set for text files, and select the unzipped file that you renamed with a 'txt' suffix;
  - In the dialogue box that lets you tell MS-Access how to parse the text file, click 'Advanced' and choose the 'specs' that apply to the particular data file (for example, ma000043)
- **Develop MS-Access query to join the geography and P43 tables.**
  - Here are the variable names that correspond to each of the 15 columns for P43 data
    -
```
     P43. SEX BY EMPLOYMENT STATUS FOR THE POPULATION
16 YEARS AND OVER [15]
       Universe: Population 16 years and over
P043001: Total:
P043002:   Male:
P043003:     In labor force:
```

```
P043004:       In Armed Forces
P043005:       Civilian:
P043006:         Employed
P043007:         Unemployed
P043008:     Not in labor force
P043009:   Female:
P043010:     In labor force:
P043011:       In Armed Forces
P043012:       Civilian:
P043013:         Employed
P043014:         Unemployed
P043015:      Not in labor force
```

    - Join the tables using the 'logrecno' column
    - Build a state+county+tract+blockgroup 12-digit block group identifier so you can join to the blockgroup map
    - Compute the percent unemployed = 100*(P043007+P0430015)/(P043005+P043012)
- **Choose appropriate summary level (150)** in order to get right counts for block groups
- **Refine and use query to pull relevant rows and columns** for block groups in all of Mass (or just for Middlesex County if we want Cambridge and its neighbors north of the Charles River).
- **Join tabular data to map** of blockgroups for Middlesex County (obtained use MIT geodata tool from Library SDE server)

This data extraction and mapping exercise is complicated because the datasets are so large and include so many variables and geographic identifiers. But it is illustrative of the issues and steps involved in (a) understanding very large and highly structured datasets, and (b) using desktop tools to find, download, and mix-n-match geometry and tabular data from different online sources.

Note that the US Census provides many online tools to obtain census data. Likewise, there are many third-party tools and CDs that repackage the data in smaller chunks, with or without maps, and sometimes in pre-processed forms (e.g., after normalizing to percent owner-occupied rather than just as the raw counts). These assorted tools fill many nitche markets. Relatively few census data users understand the data structure and raw files at the level described in these lecture notes - i.e., at the level needed to find and use any of the thousands of columns of data that are available at each level of geography..
