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

September 29, 2010

Administrative

- Homework set #1 has been posted online-- due in two weeks (Oct. 13, 2010 2 PM)
- Lab 2 has been graded, and your grades the Grading Criteria have been posted on Stellar.
- Download the PDF version of this lecture.

Census Discussion Overview - Utilizing Large Tabular Datasets (e.g., the US Census)

- Understand nature and use of large, highly structured, public datasets
    - Examine primary US Census Data (SF3=Summary File 3) at the block group level
    - Appreciate differences between Census SF3 CDs and third-party census extracts
        - Voluminous 'free' raw datasets for entire country
        - Pre-packaged 'expensive' extracts for county or metropolitan area
    - Learn how to manipulate census data in MS-Access and ArcGIS
- Understand key aspects of US Census Data
    - What is it and why do we care?
    - How are the data 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

Example - Thematic map of Income (viz., median earnings)

- **Isn't this easy?** - we did thematic map of income on day 1!
    - Yes, if desired variable is already in attribute table of map
    - There are thousands of variable in the 'long form' census
        - Which combinations zero in on **useful indicator**
        - Drilling down and combining data are often needed

    - What do we mean by 'income'
        - Household, personal??
        - Earned income, all income??
        - At what scale? state, county, city, tract, block group??
        - For what time period? weekly, yearly, part-time??
- Example: median 1999 personal earnings from the 2000 US Census
    - Variable P85 (among the hundreds of census variables and thousands of columns)
    - P85 records "Median earnings in 1999 dollars by sex for the population 16 years and over with earnings"
        - 'earnings' includes wages, salaries, and net self-employment income (but not entitlements)
        - Note differences between 'earnings' and 'income'
    - The P85 table has three columns:
        - P085001 = total (for **universe** of population 16+ years old with earnings)
        - P085002 = male
        - P085003 = female
- Use MS-Access database in class locker: M:\data\census2k\hw2_sf3_lite.mdb [Beware: 46 MB! Copy to local drive before using.]
    - Two of the 70+ raw US census files (for Massachusetts) have already been loaded
        - Table **Ma00007** - census variables P68 through P91
        - Table **Mageo** - the geographic identifiers for census tracts, block groups, counties, etc.
        - Table **blkgrp2t** - the cross reference table for block groups and towns in/around Cambridge
        - Query **mageo150** - combines state+county+tract+blockgroup to match mappable block group identifiers
            - 'Expression builder' is a powerful but hidden MS-Access tool!
    - Determine the median earnings for Cambridge blockgroups
        - Understand census data structure and use of raw data
        - Examine ER diagram of relationships among the tables used in this query
        - Illustrate SQL query development in MS-Access
            - mageo150 query pulls out block group level data and builds STCNTRBG key for ArcMap join
            - build new query to add income (**P085001**) column to mageo150
            - Save query and 'make table' to have results available in both forms
- Map median earnings for Cambridge block groups
    - Bring MS-Access query results into ArcMap
        - Add *.MDB database via filesystem - okay, but must be a 'table' and database is locked until ArcMap is closed
        - Add *.MDB via ODBC database connection - sees queries but truncates to integer
    - Create thematic map
- Examine Technical Documentation for the SF3 Census data
    - Online site at US Census: http://www.census.gov/prod/cen2000/doc/sf3.pdf
    - Copy in class locker (for faster access): http://mit.edu/11.520/data/census2k/sf3.pdf
    - Learn how to identify variables of interest and find them in the data tables

What Is the Census 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 (although now changing to a rolling census)
- Attempts an actual count of entire population categorized by various criteria
- The only source for spatially detailed demographic data with a a consistent coast-to-coast data structure
- 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

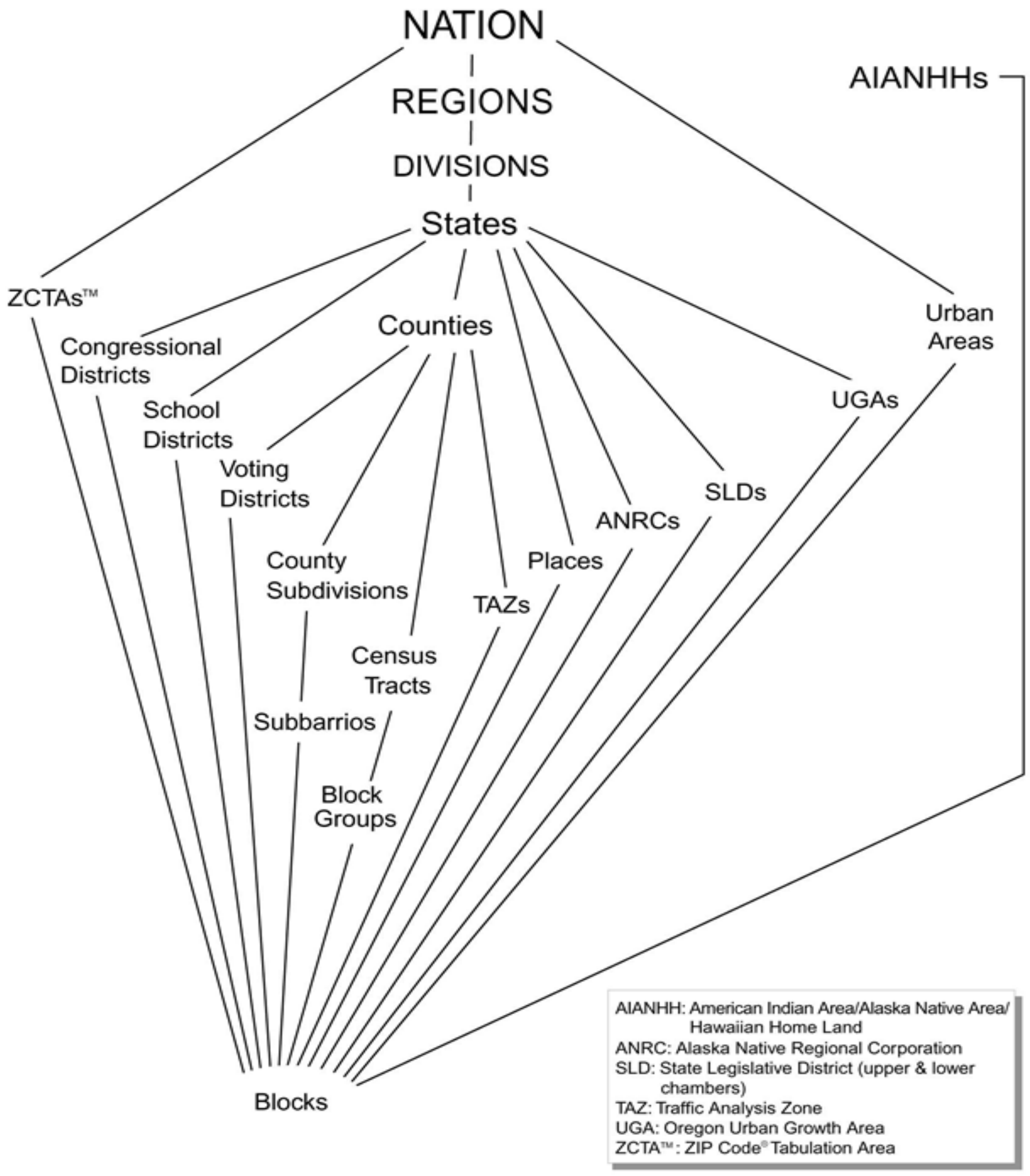
- **Why We Need to Know the Two Components**
  - Accuracy of the data varies and counts differ (Why?)
  - It helps us to understand how the data are organized in Summary Files (SFs)

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)


NATION
REGIONS
DIVISIONS
States
AIANHHs
ZCTAs™
Congressional Districts
School Districts
Voting Districts
Counties
County Subdivisions
Census Tracts
Subbarrios
Block Groups
TAZs
Places
ANRCs
SLDs
UGAs
Urban Areas
Blocks
AIANHH: American Indian Area/Alaska Native Area/ Hawaiian Home Land
ANRC: Alaska Native Regional Corporation
SLD: State Legislative District (upper & lower chambers)
TAZ: Traffic Analysis Zone
UGA: Oregon Urban Growth Area
ZCTA™: ZIP Code® Tabulation Area

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

Summary Level	Geographic Unit
010	United States

020	Region: Northeast (NE), Midwest (MW), South (S) and West (W) Regions
030	Division: Northeast Region: New England, Mid Atlantic Midwest Region: East North Central, West North Central South Region: South Atlantic, East South Central, West South Central West Region: 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**

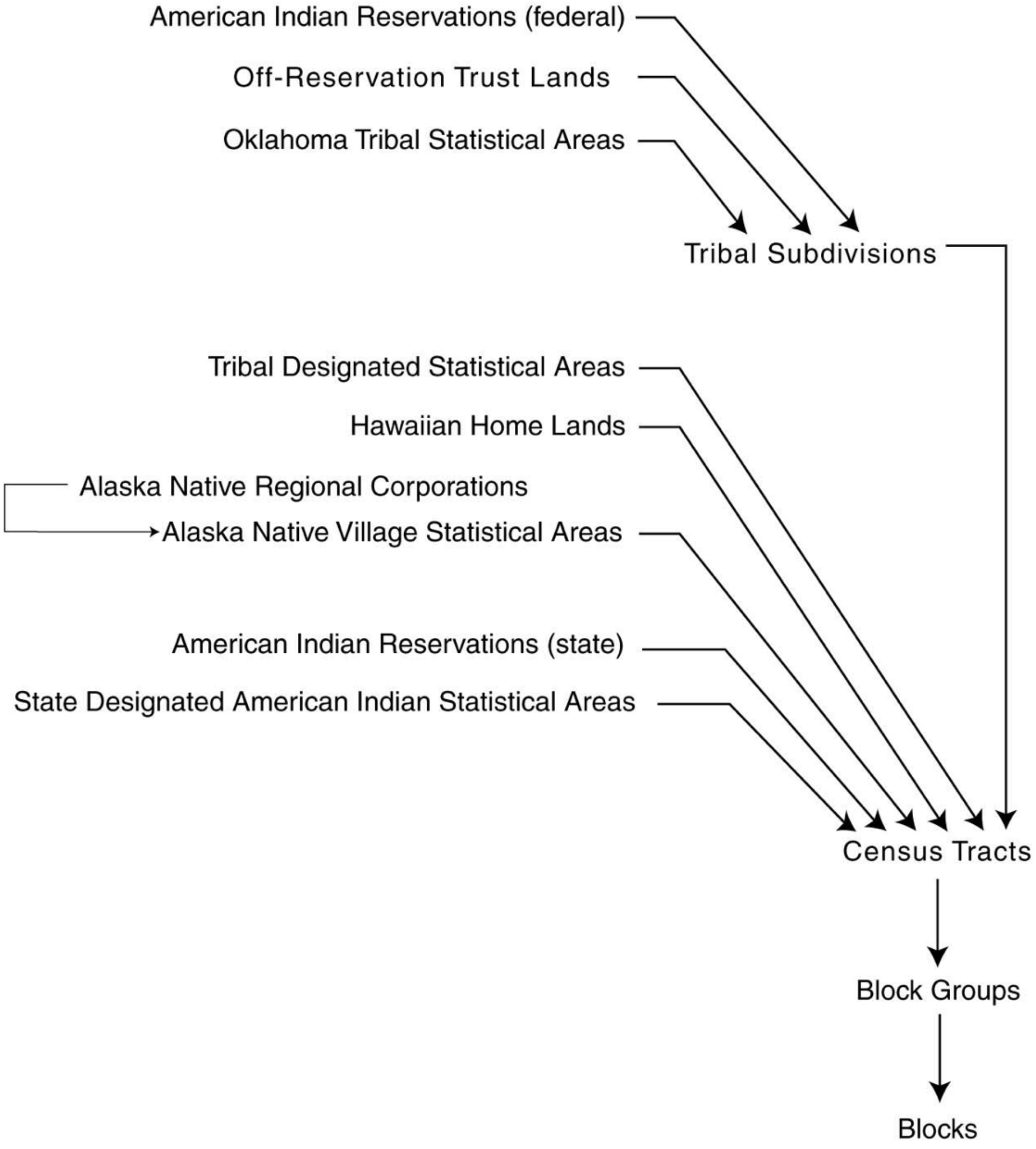
Summary Level	**Geographic Unit**
040	State (includes Washington, D.C. & Puerto Rico)
050	County
140	Census Tract
150	Block Group

- **Supplemental Geographic Areas**

Summary Level	**Geographic Unit**
400	Urbanized Areas
300	Metropolitan Areas (MSAs, CMSAs)

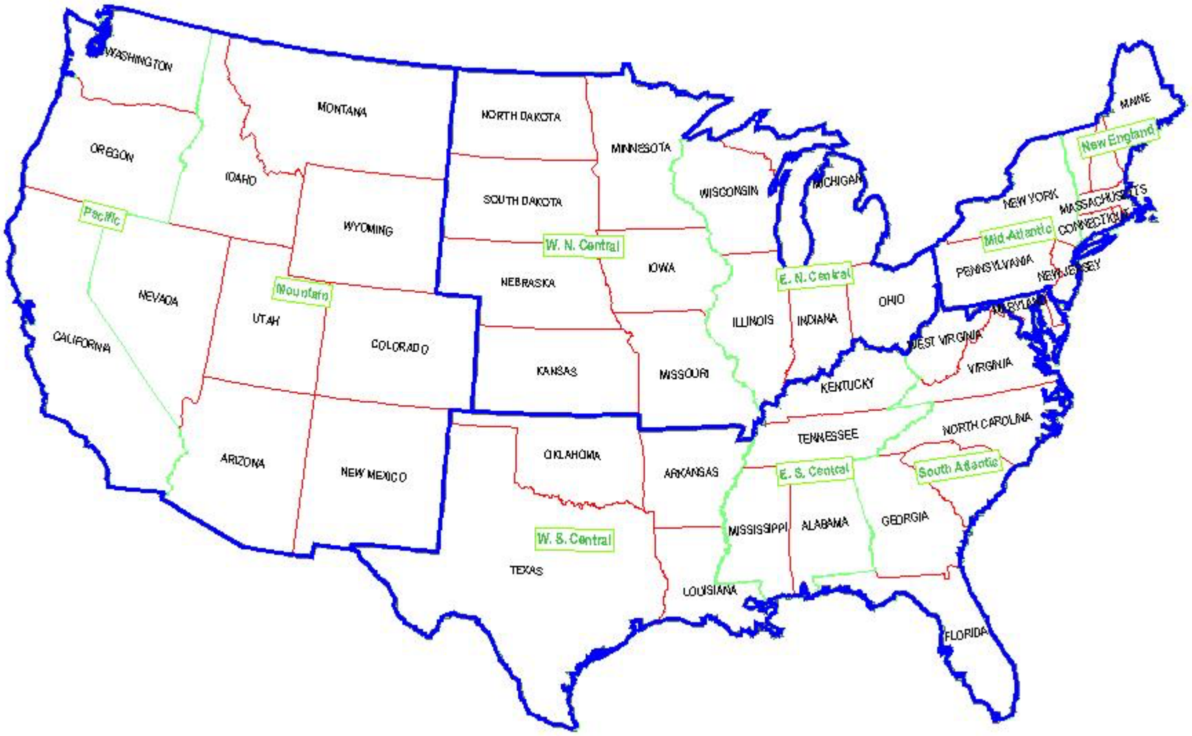
200	American Indian and Alaska Native areas
800	ZIP codes

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)

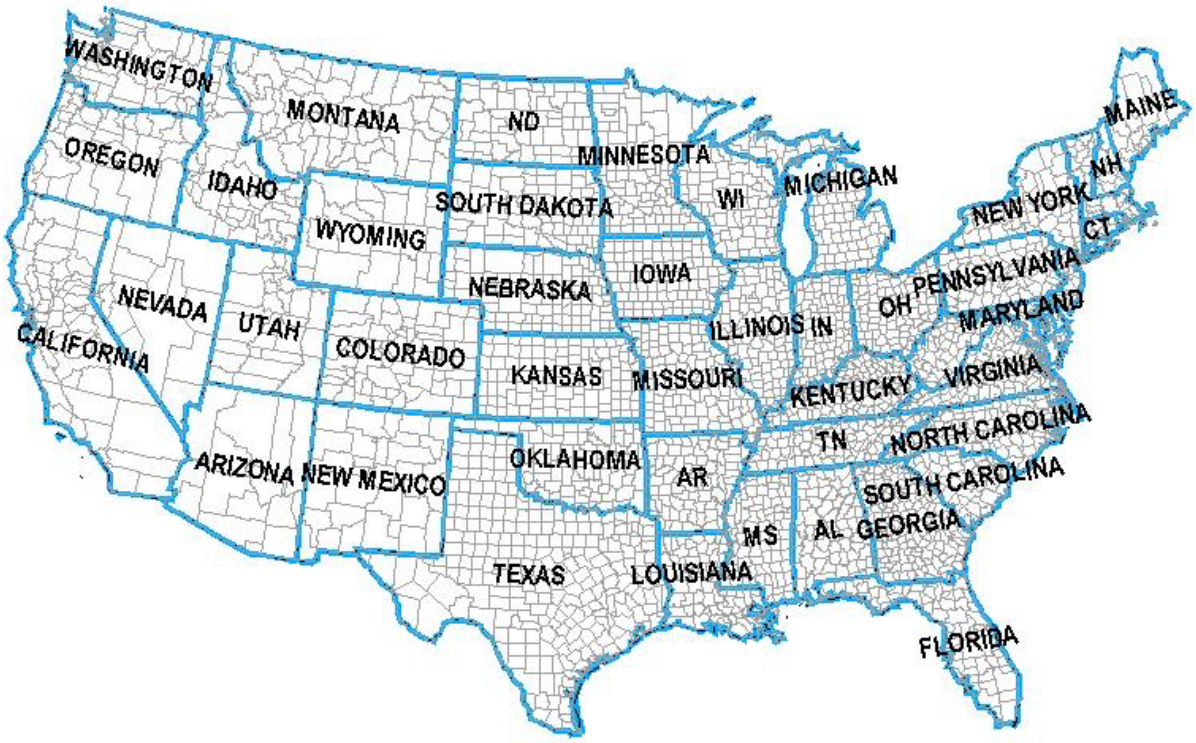

American Indian Reservations (federal)
Off-Reservation Trust Lands
Oklahoma Tribal Statistical Areas
Tribal Subdivisions
Tribal Designated Statistical Areas
Hawaiian Home Lands
Alaska Native Regional Corporations
Alaska Native Village Statistical Areas
American Indian Reservations (state)
State Designated American Indian Statistical Areas
Census Tracts
Block Groups
Blocks

- **A Visual Look at Census Geography**

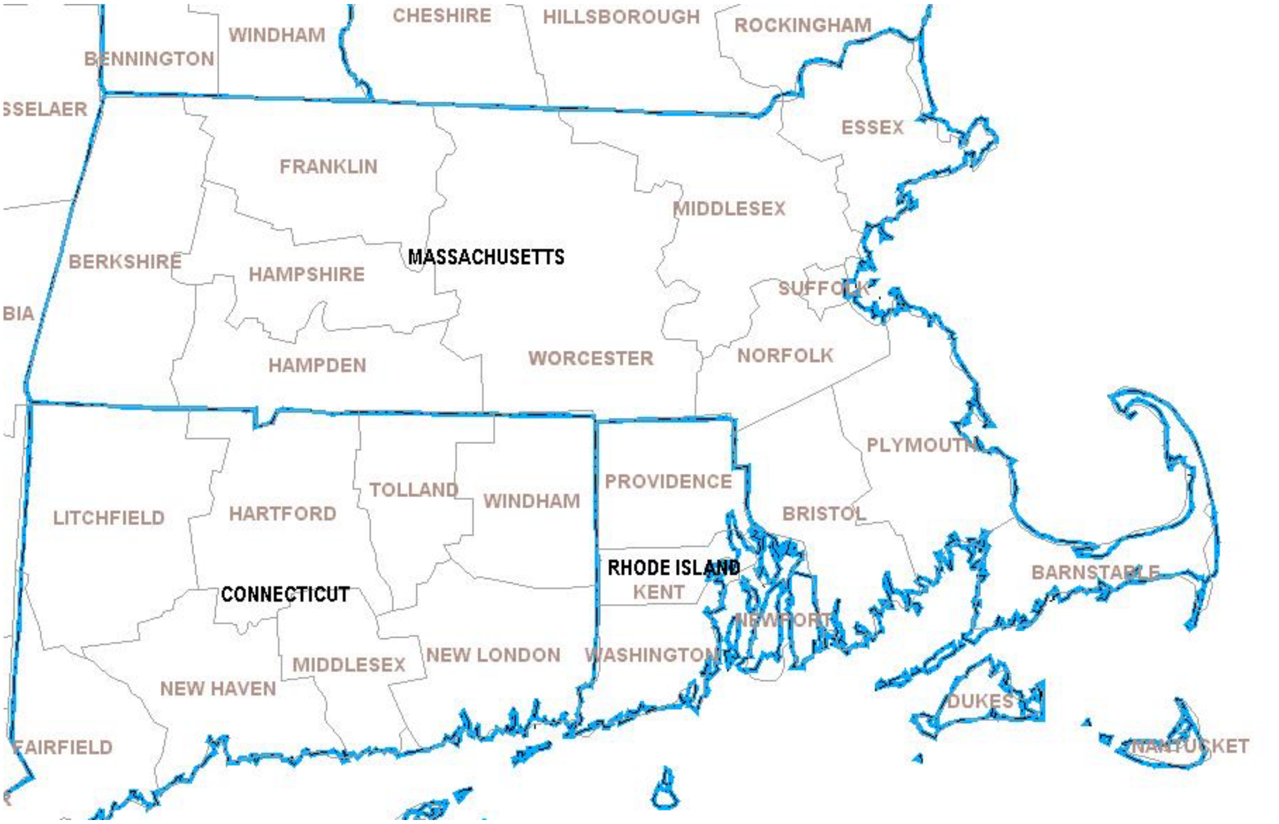
- **Continental United States (Regions in blue; Divisions in green; States in brown)**



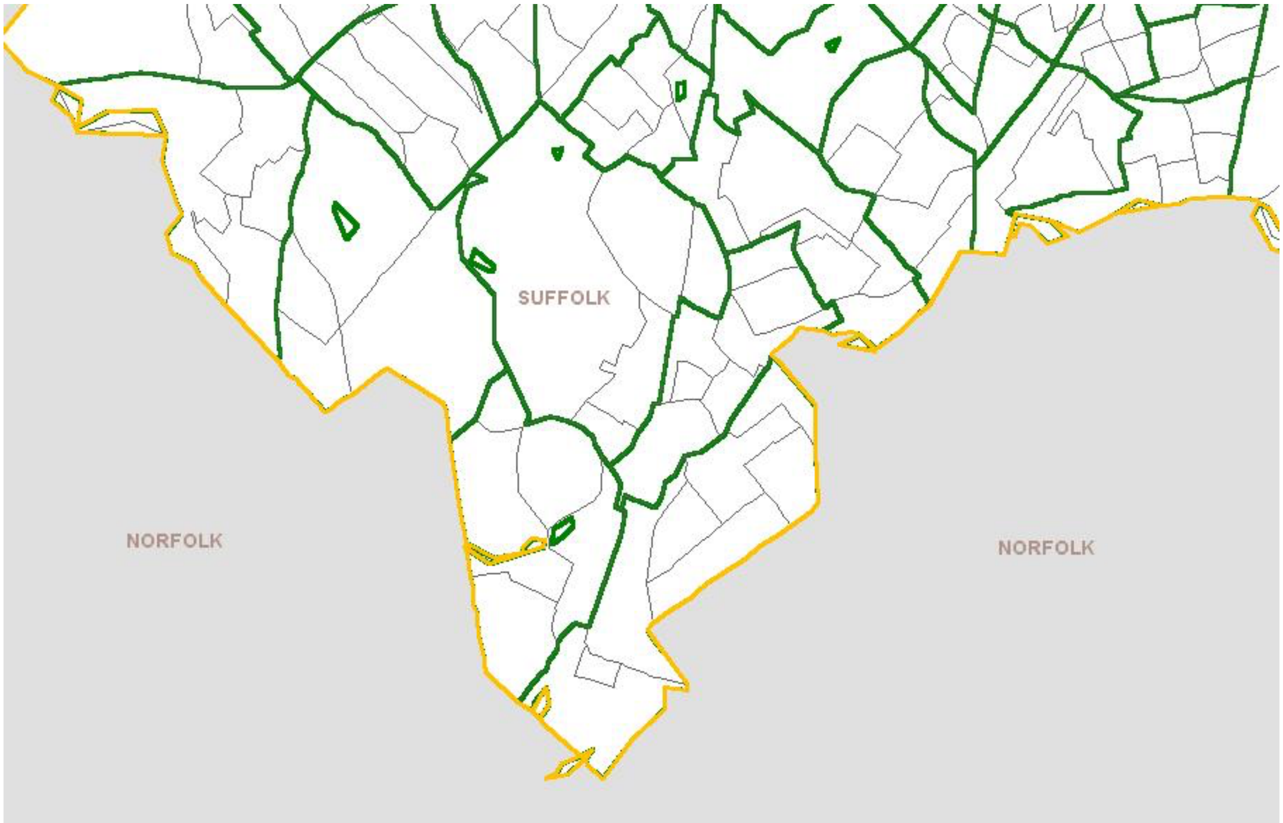
- **Counties**



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



- **Tracts (green lines) and Block Groups (gray 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
    - E.g., in SF1 P0020001= FAMILIES but in SF3: P0020001 = 100-PERCENT COUNT OF PERSONS
  - 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
    - Census TIGER 2000 Linefiles 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. American FactFinder provides a forms-based online interface to many US Census datasets including SF1 and SF3. The FactFinder website is convenient when you want data for a single census tract or a small number of areas. It is also convenient when you want a few percentages (such as percent owner-occupied) that would otherwise require downloading the numerator and denominator needed for your own calculation. If you need to download many variables or data for many areas, you may be better off accessing the core SF1 and SF3 datasets described above via the following links. (The GIS lab in the Rotch Library has many Census CDs and other third-party tools that may also be helpful. Later in the semester, we will also use the MIT Library's online geodata repository that contains direct ArcGIS access to many useful datasets including some US Census data.)

The Census Bureau distributed the 1990 Census files as DBF files on CD-ROMs. The Census Bureau has posted the contents of many 1990 CD-ROMs online. These are available via HTTP and FTP. Also,

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.

- **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

1990

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

> Note that 1990 Census CDs are also available for borrowing from the MIT Rotch Library.

2000 Census

- American FactFinder
- Public Law 94-171 (PL 94-171)
  - 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

- Commercial firms often repackage US census data
  - ESRI sample data (and online geography network) contain common census variables
  - Rotch Library has Geolytics CDs with convenient census datasets including 1970-2000 data that has been adjusted to reflect 2000 census tract boundary files
- Data Release Dates
- Subjects Areas of Questions Asked (lists first US census in which subject areas were first included - helpful when contemplating longitudinal studies)

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

- Here are detailed notes on determining unemployment rates from the raw US Census tables - including importing the raw text files into MS-Access (Note: All these steps are **NOT** needed for class and homework since we have already downloaded and formatted key Mass datasets. The extra steps preceeded by '###' are included here to clarify the complete process in case you have need for some other project to obtain other census tables besides the one used in our exercises.)
- **How should we measure unemployment rate**: Census definition is: " 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'
  - Open the MS-Access database containing the template (it is called SF3.mdb in the class locker) and 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; (For the 2007+ version of Access, use the External-Data tab and then the 'text file' choice in the 'import' section.)
  - 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+P0430014)/(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 only want Cambridge and its neighbors north of the Charles River (all of which are in Middlesex County).
- **Join tabular data to map** of blockgroups for Middlesex County (obtained use MIT geodata tool from Library SDE server)

> The four steps above that are marked with ' **#**' are **not needed** for the class exercises since we have already built an MS-Access database with the Census variables needed for the lab and homework exercises.

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..

The section of these notes entitled "Introduction to the U.S. Census of Population and Housing" is adapted from a Microsoft PowerPoint presentation originally
created by Prof. Qing Shen for 11.208 on January 21, 1997.
Augmented and modified 1999-2010 by Thomas H. Grayson, Anne Kinsella Thompson, Sarah Williams, Xiongjiu Liao, Joe Ferreira, and Shan Jiang.

Last modified: 29 September 2010 [shanjang]

Back to the 11.520 Home Page. - Back to the CRON Home Page.