Data Source Evaluation

presented by:
Tim Haithcoat
University of Missouri
Columbia
Socio-Economic Data

- Data about humans, human activities, and the space and/or structures used to conduct human activities
- Specific classes include:
  - Demographics (age, sex, ethnicity, marital status, education)
  - Housing (quality, cost)
  - Migration
  - Transportation
  - Economics (personal income, employment, occupations, industry)
  - Retailing (customer locations, store sites, mailing lists)
Disaggregated Data

- Data about individuals or single entities, for example:
  - A person’s age, sex, level of education, income, occupation, etc.
  - Gross sales, number of employees, profit, etc. for a retail store
  - Registration number and type for a single vehicle
Aggregated Data

• Describing a group of observations with the grouping made on defined criterion

Geographical data often grouped by spatial units
  • e.g. census tract, traffic zone

Aggregation can also be by time interval
  • e.g. number of persons leaving area in 5 years

Also by socio-economic grouping
  • e.g. persons aged 5-14 years

Examples of aggregated data:
  • Number of persons, average income, median housing value for a census tract
  • Number of commute trips & average trip length from a suburb to a central business district
<table>
<thead>
<tr>
<th>Cross-Sectional Data</th>
<th>Longitudinal Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gives information on many areas from the same single slice or interval of time</td>
<td>Gives information on one or more areas for a series of times</td>
</tr>
<tr>
<td>Examples:</td>
<td>Examples:</td>
</tr>
<tr>
<td>Average income in census tracts of Los Angeles for 1995</td>
<td>Average income for state of New York from 1970-1988 by year</td>
</tr>
<tr>
<td>Numbers migrating out of each state in the period of 1981-85</td>
<td></td>
</tr>
</tbody>
</table>
Sources of Economic Data for GIS (1 of 3)

• **Field surveys**
  
  Data used in marketing is gathered by door-to-door or street interviews
  
  Require careful sampling design
  
  • How to obtain a representative sample
  
  • How to avoid bias toward certain groups in street interviews

• **Government statistics**
  
  Statistics collected and reported by government as part of required activities, e.g. Bureau of the Census
  
  Usually based on entire population, except sampling is used for some Census questions
Sources of Economic Data for GIS (2 of 3)

• **Government administrative records**
  Records are collected by government as part of administrative functions
  • Examples: tax records, auto registrations, property taxes
  These are useful sources of data provided confidentiality can be preserved
  Usually available only to government or for research purposes

• **Secondary data collected by another group, often for different purposes**
  The original mandated purpose of the Census was to provide data for congressional districting
Sources of Economic Data for GIS (3 of 3)

- **Private sector companies**
  
  Retailers and direct-mail companies are major clients for these companies.
  
  Includes data originally from census augmented from other sources and surveys.
  
  Data can be customized for clients (special sets of variables, special geographical coverage or aggregation).
  
  Customizing justifies costs, which are often higher than for “raw” census data.
For use in GIS, socio-economic statistics are of little use without associated “geography”, the term often used to describe locational data

- e.g. data on census tracts must be supported by digital information on locations of census tract boundaries

- Geography also allows data to be aggregated geographically

  - e.g. by merging data on individual cities into metropolitan regions
“Geography” (2 of 3)

• Thus, many suppliers of socio-economic data also supply digitized geography of reporting zones

• Boundaries of many standard types of reporting zones change from time to time
  
  e.g. changes occur occasionally in county boundaries

  e.g. enumeration districts are redefined for each census

  Difficult to assemble longitudinal data for such units due to changing geography
“Geography” (3 of 3)

• Data is often needed for one set of reporting zones, only available for another set
  e.g. data available for census tracts, required for school districts which do not follow same boundaries
  Such problems of cross-area estimation are facilitated by GIS technology

• These problems are often grouped into the area of modifiable area problems (MAP)
  Considerable effort has been expended recently to develop statistically sound techniques to deal with these problems
## Issues in using Secondary Socio-economic Data

### Cost
- Usually less expensive than field surveys
- Large expenditures by government agencies on data collection (e.g. US Census) are indirect subsidies to users, who often pay much less than real cost of data

### Data Quality
- Major difficult is undercounting - census and other social surveys tend to miss certain groups, leading to bias in results
- Undercounting in US Census may be as high at 25% for some groups

### Documentation
- Quality of documentation, supporting information (e.g. maps) is usually high for data collected by government

### Currency
- Social data changes rapidly, can be quickly out of date (e.g. births, deaths, migration, changing economy)
- Competitive edge in retailing depends on having current data
- US census only every 10 years, so data may be 10 years old
- Often have to estimate current or future patterns based on old data
Issues in using Secondary Socio-economic Data

**Aggregation**
- Data available with suitable level of spatial, temporal aggregation?
  - e.g. study to change elementary school district boundary will require data at resolution of city blocks or higher
  - e.g. location for gas station will require city block level data, for shopping mall much lower resolution (greater aggregation of data) is adequate

**Data Conversion**
- Conversion steps may be necessary to make data useful in GIS
  - e.g. format, type of data may be incompatible

**Accuracy of Location**
- Census locates people by place of residence - “night-time” census
- “daytime” data would show locations during the data (place of work, school, etc.) but is generally not available from standard sources
- Medical records often locate individuals by place of treatment (hospital), not residence or workplace
  - e.g. consider implications for detecting exposure to cancer-causing agents
Sources of Socio-Economic Data (1 of 4)

**Population Census**
- Questions on age, sex, income education, ethnicity, migration, housing quality, etc.
- Summary of statistics used in research, planning, market research, available at high level of geographic resolution in many countries.

**Economic Census**
- Enumeration & tabulation of business activity is conducted by the US Census Bureau in years ending in 2 and 7.
- Detailed information on classes on industry.
- Low level of geographic resolution (i.e. large reporting zones).
- Data collected in many countries through annual, quarterly or monthly returns of information from companies.
Sources of Socio-Economic Data (2 of 4)

**Agricultural Census**
- Annual data on crops, yields, livestock, etc.
- More extensive periodic surveys of farm economy
- Available in spatially disaggregate form to e.g. County level in US

**Labor Force Statistics**
- Enumeration of employment, unemployment
- Produced from periodic (e.g. monthly) sample surveys of workforce
- Other special-purpose surveys often combined with regular labor force survey - e.g. household expenditures, recreation activities
- Often available for small areas, e.g. parts of city
### Sources of Socio-Economic Data (3 of 4)

#### Administrative Records
- Vehicle registrations, tax returns, etc.
- Useful for various marketing, research purposes
- Based on 100% sample so can be disaggregated spatially
  - However, disaggregation causes problems over confidentiality of records

#### Land Records
- Record of land parcel description, ownership and value for taxation purposes
- Updated on a regular basis (e.g. annually) by municipality or county government
- Also used for land use planning
- Source of current demographic information in some countries/states (i.e. local census)
Sources of Socio-Economic Data (4 of 4)

Transportation and Infrastructure Inventories

• Planning, management and maintenance of facilities
• Includes roads and streets, power lines, gas lines, water, sewer lines
• Collected by local utilities, responsible government departments
• valuable to a variety of users
  e.g. construction companies needing information on buried pipes
  e.g. emergency management departments needing data on hazardous facilities
• Compiling agency often sees a substantial market for such data which can offset costs of collection
US Census of Population & Housing

• Process of taking the census
  
  Purpose is to enumerate the population for redefining election districts
  
  Taken every ten years (1980, 1990, etc.)
  
  April 1st is census day, although complete enumeration takes a “few” weeks
  
  Most households receive forms in mail, some require visit by enumerator
1990 US Census Content
Items collected at every household ("complete-count items"):  

**Population**  
- Household relationship  
- Sex  
- Race  
- Age  
- Marital status  
- Hispanic origin

**Housing**  
- Number of units in structure  
- Number of rooms in unit  
- Tenure (owned or rented)  
- Value of home or monthly rent  
- Congregate housing (room and board)  
- Vacancy characteristics
1990 US Census Content
Additional items collected at sample households:

**Population**
Social Characteristics:
- Veteran status
- Education - enrollment and attainment
- Place of birth, citizenship and year of entry to US
- Ancestry
- Language spoken at home
- Migration (residence since 1985)
- Disability
- Fertility

**Economic Characteristics:**
- Income in 1989
- Labor force
- Occupation, industry & class of worker
- Place of work and journey to work
- Work experience in 1989
- Year last worked

**Housing**
- Year moved into residence
- Number of bedrooms
- Plumbing & kitchen facilities
- Telephone in unit
- Vehicles available
- Heating fuel
- Source of water and method of sewage disposal
- Year structure built
- Condominium status
- Farm residence
- Shelter costs, including utilities
Process of Census Returns

- Automated encoding to digital form
- Automated editing to correct obvious inconsistencies
- Some missing items can be assigned automatically using simple rules
- Other missing items are assigned based on probabilities
- Data assembled into master database
- Sample surveys processed to produce statistical summaries
Geographic Referencing

• Initially returns are identified by street address
• Address is converted into geographic location using a digital referencing system
  For the 1980 census, DIME (Dual Independent Map Encoding) files were used for digital geographic referencing or urbanized portions of the US
  For the 1990 census, TIGER files covering every county was used
• TIGER has a major impact on GIS databases
Census Reporting Zones

- Range from blocks to states
- Hierarchy of census areas, 1990 (next slide)
- 1990 census units (following next slide)
- As noted previously, the geographic boundaries and definitions of these areas may change from one census to the next
Hierarchy of Census Areas, 1990
(Values in brackets indicate average population)
1990 US Census Units
Political, Governmental & Administrative Units

- States, district of Columbia and state equivalents: Puerto Rico, Guam, the Virgin Islands, American Samoa, Palau and the Commonwealth of the Northern Marina Islands
- Congressional districts
- Voting districts
- Counties
- Minor Civil Divisions (MCD’s) - the primary political and/or administrative subdivisions of a county, “townships” in many states
- Incorporated places
- American Indian reservations and trust lands
- Alaska Native Regional Corporations
1990 US Census Units
Statistical Units: Defined by the Census Bureau (1 of 2)

- **Regions and Divisions:** the US is divided into four regions each with 2-3 divisions

- **Metropolitan Statistical Area (MSA), formerly SMSA:** consists of one or more counties including a large population nucleus and nearby communities that have a high degree of interaction

- **Urbanized areas (UA):** defined by population and density, population more than 50,000 and density > 1,000 mi$^2$

- **“Urban/Rural”:** all people in UAs and places of >2,500 population are “urban”, all others are “rural”
1990 US Census Units
Statistical Units: Defined by the Census Bureau (2 of 2)

- **Census tracts**: small, relatively permanent areas partitioning large cities and some counties, average 4,000 population
- **Block Numbering Area (BNA)**: equivalent to Census Tract in rural areas
- **Enumeration Districts (ED)**: not used in 1990
- **Block Groups (BG)**: groups of blocks within Census Tracts averaging 1000 population, replace EDs
- **Block**: smallest Census area, contain about 70 people, in urban areas may be a single city block. (note: 1990 census was first in which data was available at this level for the entire nation.)
Availability of Census Data (1 of 2)

- Tabulation of statistics by reporting zones (e.g. population by county, population by age by county)
- Crosstabulation, e.g. population by age and sex by county
- Special tabulations, e.g. for unusual combinations of characteristics, or for unusual or custom reporting zones
- Number of possible tabulations and crosstabulations is infinite, volume of census produces vastly exceeds volume of data collected
Availability of Census Data (2 of 2)

- Alternative formats for products
  - Printed reports
  - Magnetic media - tapes, disks
  - Microfiche, microfilm, now CDs

- Sources of census data
  - State data centers distribute Census data
  - Private firms repackage and customize data, produce custom reports (e.g. tabulation of population by distance form proposed mall location)

- Geography products available
  - Base maps showing reporting zones
  - Atlases produced for urban areas
  - Digital products - boundary files, TIGER
1990 US Census Products

- Printed reports
- Computer tapes
  - Summary tape files
  - Subject summary tape files
  - Public use microdata sample files
- Other media
  - Online information systems
  - CD-ROM
- Microfiche
- Geographic publications

- Maps
  - County block maps
  - County subdivision maps
  - Census tract/BNA outline maps
- Custom data tabulations
  - User defined area tabulations
  - Special tabulations
- Machine readable geographic files
  - TIGER
Topologically Integrated Geographic Encoding and Referencing System
Development

• Designed to:
  Support pre-census geographic and cartographic functions in preparation for the 1990 Census
  To complete and evaluate the data collection operations of the census
  To assist in the analysis of the data as well as to produce new cartographic products

• TIGER files were created by the Bureau of the Census with the assistance of the US Geological Survey
Content

• TIGER/line files are organized by county
• They contain:
  Map features such as roads, railroads and rivers
  Census statistical area boundaries
  Political boundaries
  In metropolitan areas, address ranges and zip codes for streets
Marketing TIGER files

• Census Bureau
  1990 census versions of TIGER/Line files are available from the Census Bureau
  Pre-census files are also available on CD-ROM

• Third-Party vendors
  Many market repackaged versions of TIGER/Line files, in many cases with software which will enable users to access this data easily and quickly
  Many of these products are designed for use on micro-computers
Non-census uses for TIGER

• TIGER files are valuable for other purposes
e.g. locating customers from address lists
e.g. planning vehicle routes through city streets, for parcel delivery, cab dispatching
For these purposes TIGER files need to be kept current at all times, but Bureau of Census only requires them to be current every 10 years
TIGER SYSTEM

TIGER SOFTWARE

TIGER FILE DATABASE

Other Data Files
- PL 94-171 FILE
- Census STF
- Land use
- Tax Parcel

GIS SYSTEM

Analytical Capabilities
Screen Displays
Map Output

TIGER/Line File
Hierarchical Relationship of Geographic Entities in the TIGER/Line Files

- Nation
- States
- Counties
- Census Tracts/BNAs
- BGs
- Blocks
- Sub-MCD’s
- County Subdivisions
- VTDs
- School Districts
- CDs
- Places
- AI/ANA’s
- ANRC’s
- UAs
Land Records

• Many systems have been developed by local governments in the US to manage land, particularly in urban areas.

• In other countries there has been more effective coordination at provincial and national levels, e.g. Australia.

  Practices in different countries depend on the system of land tenure.

• The basic entity in land records systems is the land parcel, i.e. the basic unit of ownership.

• Traditionally, land records have been managed by hand using methods which often date back 200 years.

• Land records are the basis of their system of local taxation, administration, as well as transfer of ownership and subdivision.
Issues in Land Records Modernization

• Accurate land records systems require accurate base mapping at a large enough scale, e.g. 1:1,000
  Such base mapping is not normally available in the US, only the wealthiest governments can afford to create it, e.g. from air photos
  The term cadaster is used for mapping of land ownership
• The cost of building land records systems can often be recovered, at least partially, from sales of data (e.g. to utilities, real estate developers) and use in other departments
  The term multi-purpose cadaster (MPC) describes the idea of using the cadaster for many purposes
• Because land records systems are being developed independently by many different jurisdictions, there is little standardization of approach, software, etc.