NAACCR Webinar Series:
GIS Applications

May 6, 2010
9:00 am, Eastern Daylight Time
-- repeats at --
2:00 pm, Eastern Daylight Time

Questions
• Please use the Q&A panel to submit your questions
• Send questions to “All Panelists”

Fabulous Prizes
Topics
1. Geography basics
2. Geographic Information Systems (GIS) basics
3. Geocoding
   a. Individual records
   b. Batch geocoding
4. Keyhole Markup Language (KML)
5. Geographic accessibility
6. Geographic distances & time

1. Geography Basics
   • Every point on the earth’s surface can be described in terms of a latitude and longitude
     – Northern hemisphere = + latitude
     – Western hemisphere = - longitude

   Galveston, TX
   • 29 degrees latitude
   • -95 degrees longitude
Degrees

• Degrees/minutes/seconds
  – Example 30°30’0”
  • 60 seconds in a minute
  • 60 minutes in a degree

• Decimal degrees *(NAACCR standard)*
  – Example 30.500000°

How large is a degree?

• The metric system originally defined the meter as one-tenth-millionth of the distance from the equator to the pole
  – Thus 90° of latitude=10 million meters, so 1° of latitude=111.11 km, about 69 miles
  – 1° of longitude is smaller than 1° of latitude (related by cosine function)

Latitude/longitude is customarily reported to six decimal places *(NAACCR standard)*

36.124579°, -75.034678°
Important census subdivisions (US)

- 50 states + DC + territories
  - California 36 million, Wyoming 0.5 million, Palau 20,000
- 3,143 counties or county equivalents
  - Los Angeles 10 million, Loving County, TX 60
    - These change slightly over time
- 70 thousand census tracts
  - Average 4,000 people; most 2,500-8,000
    - These change every 10 years

Important census subdivisions (US)

- 200 thousand Block Groups
  - Subdivisions of census tracts, 1-9 per tract
  - Average population 1,500
- 8 million Blocks
  - Areas bounded by roads or water
  - Most have <50 people, many have 0

Other subdivisions of space

- ZIP codes
  - Census ZCTA (Zip Code Tabulation Area)
- Cities, towns, boroughs, villages
  - (Minor Civil Divisions)
- School districts
- Health districts
- Legislative districts
- Hospital service areas
Location

- Our data reflect the location of your residence at the time you get cancer
  - It says little about lifetime exposure and even recent exposure
  - "You are where you live"

2. GIS basics

"Information systems used to manipulate, summarize, query, edit, visualize - generally, to work with information stored in computer databases." (NCGIA)

"In the strictest sense, a GIS is a computer system capable of assembling, storing, manipulating, and displaying geographically referenced information, i.e. data identified according to their locations. " (USGS)

GIS software includes several components:

- Algorithms and methods for spatial operations (e.g. adjacency), querying, editing, and creating/capturing data, and visualization (e.g. overlay)
- Database management systems (e.g. Oracle, SQL, Access)
- Open software architecture allowing customization (e.g. Visual Basic, Java, Python)
- Supplemental tools for spatial analysis, spatial/geostatistics, network analysis, geocoding, geo-rectification
How is GIS useful for Epidemiology?

• Displays and stores data at various geographic scales
  – County, census tract, neighborhood
• Provides a systematic way to explore associations between health outcomes and various risk factors (environmental, social)
• Data from various sources and geographic scales can be combined to provide estimates of exposure or area-based characteristics

How is GIS Useful for Epidemiology?

• Analyzes spatial relationships (e.g. proximity of cases from treatment facility or screening facility)
• Provides tools for creating maps for communicating results
Why Map Cancer Data?

• Public and local cancer coalitions want information about local communities
• Public health officials (state, local federal) want to understand:
  – What areas are in need of intervention and/or prevention
  – Where to allocate financial resources
  – Track progress on cancer control and prevention activities

Why Map Cancer Data?

• Identify broad spatial patterns (descriptive)
• Hypothesis generation (cause of pattern)
• Define and compare patterns (health disparities)
• Changing patterns over time (cancer control)

Mapping software

• Online or stand-alone software products mostly for descriptive mapping
• Increasingly, online mapping products/services are adding additional methods analogous to GIS
  – Annotation
  – Data creation
  – Geocoding
  – Driving distances
Mapping software

• Freely available mapping tools
  • Google Earth
  • Microsoft: Virtual Earth; MapPoint
  • ESRI Arc Explorer

Google Earth

• Virtual globe, map and geographic information program
• Provides tools for viewing, editing, creating geographic layers, geocoding, and obtaining driving directions

Google Earth

• Internal coordinate system is geographic coordinates (latitude/longitude) on the World Geodetic System of 1984 datum
• Baseline resolutions for continental U.S. of 1 m
• JavaScript API
• Free
3a. Geocoding

- Geocoding (verb): act of transforming aspatial locationally descriptive text into a valid spatial representation using a predefined process.
  - Manual or automated process
- Geocoder (noun): a set of inter-related components in the form of operations, algorithms, and data sources that produce a spatial representation for locationally descriptive text
- A geocode (noun) is a spatial representation of a descriptive locational reference.


Geocoding


Possible geocoding workflow at central registry

Geocoding process details

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Usage</th>
<th>Best/Worst output Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>The University of Southern California</td>
<td>Named place</td>
<td>County, Counts</td>
<td>Parcel-level/Non-matchable</td>
</tr>
<tr>
<td>The University of Southern California GIS Research Lab</td>
<td>Named Place</td>
<td>Screening, Disparities</td>
<td>Sub Parcel-level/Non-matchable</td>
</tr>
<tr>
<td>The northeast corner of Vermont Avenue and 36th Place</td>
<td>Relative Intersection</td>
<td>Screening, Disparities</td>
<td>Intersection-level/Non-matchable</td>
</tr>
<tr>
<td>3420 South Vermont Ave, Los Angeles, CA 90009</td>
<td>Street Address</td>
<td>Screening, Disparities</td>
<td>Building-level/Street-level</td>
</tr>
<tr>
<td>USP ZIP Code 90089-0255</td>
<td>USP ZIP Code</td>
<td>County, Counts</td>
<td>Building-level/USP ZIP Code-level</td>
</tr>
<tr>
<td>34.022351, -118.291147</td>
<td>Geographic coordinates</td>
<td>Screening, Disparities</td>
<td>Sub Parcel-level/Non-matchable</td>
</tr>
</tbody>
</table>

Data cleaning

Reference data source

- Reference dataset: geographic database containing features and address data that the geocoder uses to generate geographic outputs

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. Census Bureau’s TIGER/Line files</td>
<td>Street centerlines</td>
<td>US</td>
</tr>
<tr>
<td>NAVTEQ Streets/NAVTEQ 2010</td>
<td>Street centerlines</td>
<td>Worldwide</td>
</tr>
<tr>
<td>Los Angeles (LA) County Assessor Parcel Data (Los Angeles County Assessor 2008)</td>
<td>Vector polygon file</td>
<td>Los Angeles</td>
</tr>
<tr>
<td>US ZIP Code</td>
<td>Point</td>
<td>US</td>
</tr>
<tr>
<td>Canadian postal code</td>
<td>Point</td>
<td>Canada</td>
</tr>
</tbody>
</table>

Geocode Metadata

- Metadata: Descriptions associated with data that provide insight into its attributes.
  - Geocoder software/vendor
  - Dataset name (e.g. TIGER/Line files)
  - Dataset year, version
  - Feature-matching strategy (e.g. deterministic or probabilistic)
Geocode Metadata

- Metadata: Descriptions associated with data that provide insight into its attributes (cont).
  - Dataset type (points, lines, polygons)
  - Feature interpolator strategy (e.g., Address-range interpolation; lot interpolation)
  - Attribute matching weights
  - Date geocoded
  - Coordinate quality (NAACCR Standard)

NAACCR V11 Data Standards and Data Dictionary

<table>
<thead>
<tr>
<th>GIS Coordinate Quality Codes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>Coordinates derived from local government-maintained address points, which are based on property parcel locations, not interpolation over a street segment's address range</td>
</tr>
<tr>
<td>01</td>
<td>Coordinates assigned by Global Positioning System (GPS)</td>
</tr>
<tr>
<td>02</td>
<td>Coordinates are exact match of house number and street, and based on property parcel location</td>
</tr>
<tr>
<td>03</td>
<td>Coordinates are exact match of house number and street, interpolated over the matching street segment's address range</td>
</tr>
<tr>
<td>04</td>
<td>Coordinates are street intersections</td>
</tr>
<tr>
<td>05</td>
<td>Coordinates are at midpoint of street segments (missing or invalid building number)</td>
</tr>
<tr>
<td>06</td>
<td>Coordinates are address ZIP code+4 centroid</td>
</tr>
<tr>
<td>07</td>
<td>Coordinates are address ZIP code+2 centroid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GIS Coordinate Quality Codes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>08</td>
<td>Coordinates were obtained manually by looking up a location on a paper or electronic map</td>
</tr>
<tr>
<td>09</td>
<td>Coordinates are address 5-digit ZIP code centroid</td>
</tr>
<tr>
<td>10</td>
<td>Coordinates are point ZIP code of Post Office Box or Rural Route</td>
</tr>
<tr>
<td>11</td>
<td>Coordinates are centroids of address city (where address ZIP code is unknown or invalid, and there are multiple ZIP codes for the city)</td>
</tr>
<tr>
<td>12</td>
<td>Coordinates are centroid of county</td>
</tr>
<tr>
<td>98</td>
<td>Latitude and longitude are assigned, but coordinate quality is unknown</td>
</tr>
<tr>
<td>99</td>
<td>Latitude and longitude are not assigned, but geocoding was attempted; unable to assign coordinates based on available information</td>
</tr>
<tr>
<td>Blank</td>
<td>GIS Coordinate Quality not coded</td>
</tr>
</tbody>
</table>
Exercise 1:
Using Google Earth to Geocode Individual Cases

Geocoding exercises with individual cases – Summary

• Accurate and precise
• Time-intensive
• Not practical for a large number of cases
• What’s the best way to geocode thousands of cases all at once?

3b. Batch geocoding

• A geocoding process that operates in an automated fashion and processes more than a single record.
• Traditionally, this has been done by interpolating along a street segment.
### Geocode 5019 W Lake Rd, Wilson NY 14172

<table>
<thead>
<tr>
<th>ID</th>
<th>637913638</th>
</tr>
</thead>
<tbody>
<tr>
<td>FromLeft</td>
<td>5009</td>
</tr>
<tr>
<td>ToLeft</td>
<td>5055</td>
</tr>
<tr>
<td>FromRight</td>
<td>5008</td>
</tr>
<tr>
<td>ToRight</td>
<td>5054</td>
</tr>
<tr>
<td>Prefix</td>
<td>W</td>
</tr>
<tr>
<td>Name</td>
<td>LAKE</td>
</tr>
<tr>
<td>Type</td>
<td>RD</td>
</tr>
<tr>
<td>ZIPLeft</td>
<td>14172</td>
</tr>
<tr>
<td>ZIPRight</td>
<td>14172</td>
</tr>
</tbody>
</table>

- **StateLeft**: 36  
- **StateRight**: 36  
- **CountyLeft**: 063  
- **CountyRight**: 063  
- **TractLeft**: 0243.02  
- **TractRight**: 0243.02  
- **BlockLeft**: 1000  
- **BlockRight**: 1002  
- **FromLat**: 43.326627  
- **ToLat**: 43.326649  
- **FromLon**: -78.761419  
- **ToLon**: -78.760191

### Linear Interpolation

\[
\text{Linear Interpolation} = \frac{(\text{Max Address} - \text{address}) \times \text{range}}{(5055-5019) / 46} \times 100 = 78 \text{ meters}
\]
Probabilistic versus deterministic approaches to match to a street segment

Deterministic matching

- The deterministic approach requires an exact match on all fields being compared
- Different passes can be made that compare different subsets of fields

Geocode 28 LAURE PL, Spring Valley, NY 10977

- Pass 1: Match on number, name, type, and ZIP code
  - Result: no matches
- Pass 2: Match on number, name, and ZIP code only
  - Result: no matches
Goal: Geocode 28 LAURE PL, Spring Valley, NY 10977

- **Pass 3**: Match on number, name soundex, type, and ZIP code
  - Result: one match – 28 LAURA PL
  (Soundex is a function that considers most of the consonants and ignores the vowels in a string of text. The soundex of Laure is L600).

Geocode 28 LAURE ST, Spring Valley, NY 10977

- **Pass 1**: Match on number, name, type, and ZIP code
  - Result: no matches
- **Pass 2**: Match on number, name, and ZIP code only
  - Result: no matches
- **Pass 3**: Match on number, name soundex, type, and ZIP code
  - Result: no matches
- **Pass 4**: Match on number, name soundex, and ZIP code
  - Result: 2 matches
    - 28 LAURA PL
    - 28 LAURA LN

When there are no matches, or 2 or more equal matches, the results should be manually reviewed (the correct answer is 28 Laurel St)

**Probabilistic matching**

- Each field being compared has a weight.
- If the fields match, a score equal to the weight is assigned. If they do not match, the score is zero.
- The highest match above a certain score is the one taken.
- Can be done in a single pass, or multiple passes
Geocode 28 LAURE PL, Spring Valley, NY 10977

• Match on number, name, type, and ZIP code
  – Suppose the calculated weights are: number 6.9, name 7.5, type 1.5, ZIP code 6.2

<table>
<thead>
<tr>
<th>Address</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 LAURA PL, 10977</td>
<td>20.1 (highest)</td>
</tr>
<tr>
<td>128 LAURA PL, 10977</td>
<td>13.2</td>
</tr>
<tr>
<td>28 LAURA PL, 11213</td>
<td>13.9</td>
</tr>
<tr>
<td>43 RAMSEY PL, 12208</td>
<td>1.5</td>
</tr>
</tbody>
</table>

• Usually, weights are specific to each address
  – Yznaga has a higher weight than Maple because it is less common

• However, some geocoding software gives the same weight to every field

Probabilistic matching results in more matches...both correct and incorrect
Reasons For Non Matches

• Abbreviated & Misspelled addresses

<table>
<thead>
<tr>
<th>Street Name Submitted to Registry</th>
<th>Correct Street Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bevrbrk</td>
<td>Beaverbrook</td>
</tr>
<tr>
<td>12 A Hadon Rd</td>
<td>120 Haddon, Apartment A</td>
</tr>
<tr>
<td>122 Stikhom Road</td>
<td>122 Stockholm Road</td>
</tr>
<tr>
<td>400 Prin-Lawre Road</td>
<td>400 Princeton Lawrenceville Road</td>
</tr>
<tr>
<td>13 A Green Street</td>
<td>13 Green Street, Apt A</td>
</tr>
<tr>
<td>Unit B, 1200 Comm Ave.</td>
<td>1200 Commonwealth Ave</td>
</tr>
</tbody>
</table>

• Apartment numbers or street addresses included in the building field (unit designations belong in a separate field).
• The address is missing one or more components (building number is missing, street type or direction is missing or incorrect).

Reasons For Non Matches

• Post office box is given rather than a street address.

• Rural route address is given.

• A building or facility name is given instead of an address (e.g., just the nursing home or apartment tower name is given).

• Missing, incorrect, retired or new Zip Codes (US Postal Bulletin reports changes)

[Image: USPS Postal Bulletin, February, 1995]


Reasons for Non Matches

• Street name changes and missing streets from reference database (e.g., new developments)
Geocoding software/services

<table>
<thead>
<tr>
<th>Name</th>
<th>Free Limit/Cost</th>
<th>API Reference Data</th>
<th>Type</th>
<th>Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>TeleAtlas Locator</td>
<td>500 free, 15,000-$410, 50,000-$1,535</td>
<td>yes TeleAtlas</td>
<td>Online encrypted</td>
<td></td>
</tr>
<tr>
<td>ThinkGeo</td>
<td>setup fee: $2,500.00, monthly: $200.00-$15,000</td>
<td>yes Tiger 2009</td>
<td>Online or PC-Based encrypted</td>
<td></td>
</tr>
<tr>
<td>Spatial Source</td>
<td>500 free, 5,01/geocode after that; min charge of $250</td>
<td>yes</td>
<td>Online encrypted</td>
<td></td>
</tr>
<tr>
<td>Bing</td>
<td>cost of ArcGIS license $1500</td>
<td>yes G-NAF</td>
<td>Online</td>
<td></td>
</tr>
<tr>
<td>TerraPages</td>
<td>no $250</td>
<td>Online or PC-Based</td>
<td>encrypted</td>
<td></td>
</tr>
<tr>
<td>MapTools</td>
<td>$3000; $300 updates</td>
<td>no Tiger</td>
<td>PC-Based</td>
<td>local</td>
</tr>
<tr>
<td>Geobase</td>
<td>no</td>
<td>no Tiger or GDT</td>
<td>PC-Based</td>
<td>local</td>
</tr>
<tr>
<td>Yahoo</td>
<td>50,000/day</td>
<td>yes Navteq</td>
<td>unsecure</td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>15,000/day</td>
<td>yes TeleAtlas</td>
<td>GDT</td>
<td>unsecure</td>
</tr>
<tr>
<td>USC Geocoder</td>
<td>2500 batches; $0.002 per geocode</td>
<td>yes Tiger 2000</td>
<td>Online encrypted</td>
<td></td>
</tr>
<tr>
<td>Maps Online</td>
<td>no</td>
<td>no Tiger or GDT</td>
<td>PC-Based</td>
<td>local</td>
</tr>
<tr>
<td>Geocoder.us</td>
<td>50,000/day</td>
<td>yes TIGER</td>
<td>unsecure</td>
<td></td>
</tr>
<tr>
<td>Adapted from:</td>
<td><a href="https://webgis.usc.edu/Services/Geocode/About/GeocoderList.aspx">https://webgis.usc.edu/Services/Geocode/About/GeocoderList.aspx</a></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exercise 2: Batch Geocoding
Batch Geocoding

- Summary
  - Knowledge of record linkage theory is helpful
  - Before selecting a vendor evaluate geocode match accuracy and review available metadata
  - Some software and vendor solutions are expensive
  - Some cases will still require individual clerical review

4. Keyhole Markup Language (KML)

- An XML-based open-standard markup language for displaying geographic data and annotation on Internet-based Earth browsers such as Google Earth.
- KML is an open standard officially named the OpenGIS® KML Encoding Standard. It is maintained by the Open Geospatial Consortium, Inc.
- KML specifies a set of features (placemarks, images, polygons, 3D models, textual descriptions) in an Earth browser

KML example (placemark)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
  <Placemark>
    <name>Simple placemark KML example</name>
    <description>
      Attached to the ground. Intelligently places itself at the height of the underlying terrain.
    </description>
    <Point>
      <coordinates>-74.688705, 40.2552900</coordinates>
    </Point>
  </Placemark>
</kml>
```
Exporting geographic data from SAS to KML

Syntax:

```sas
filename KMLOPT 'C:\GISApplications_POI\import\KML_TO_SAS.map';
libname KMLOPT 'C:\GISApplications_POI\import\KML_TO_SAS.map';
```
Exercise 3: Creating and Editing Geographic data in Google Earth

5. Geographic Accessibility

- Access: Ability to obtain appropriate health services when required

- Components of Access (Ricketts & Savitz, 1994):
  - Availability
  - Accommodation
  - Affordability
  - Acceptability
  - Accessibility

Accessibility

- Relative ease by which health services can be reached, taking account of resources (e.g. transportation, physicians) and travel time, distance and cost.

- Types of Accessibility:
  - Revealed (or Realized)
    - Actual use of a service
  - Potential
    - Probable utilization of a service
Accessibility examples in cancer research

- Revealed:
  - For cancer patients, increased travel distance/time has been shown to be associated with:
    - A greater risk of presenting with advanced stage cancer
    - Decreased utilization of breast-conserving therapy
    - Lower enrollment in clinical trials

- Potential:
  - Hayanda et al. (2009) found that each percentage point increase in the African American population within a county was associated with a decrease in the number of colorectal surgeons and gastroenterologists.


Accessibility examples in cancer research

- Potential:
  - Omega et al. (2008) found:
    - For the total continental U.S. population, the median travel time to the nearest NCI Cancer Center is 78 minutes.
    - Native Americans and nonurban residents have relatively longer travel times to NCI Cancer Center, academic care, and specialized cancer care.
    - Travel times of 1 hour were estimated for 45.2% of the population to the nearest NCI Cancer Center, 69.4% to the nearest academic-based care, and 91.8% to any specialized cancer care.

Commonly used measures of geographic accessibility

1. Provider-to-population ratios:
   - Physicians per # of population
   - Does not include distance or time
   - Does not account for patient border crossing (e.g. county)

Commonly used measures of geographic accessibility

2. Travel distance or time to the closest service
3. Number of services within n miles or minutes
4. Average travel distances or times to all or n closest services
5. Floating catchment or Gravity-based methods
   - Account for supply & demand

6. Distance Measures
   - Great Circle Distance (GCD)
     - “Shortest distance between any two points on the surface of a sphere measured along a path on the surface of the sphere.”
     - More accurate than straight-line (linear) distance
     - Prior to SAS 9.2 Haversine formula used to calculate GCD
     - 9.2 uses Vincenty’s formula (more precise than Haversine formula)
Distance Measures

- Driving Distance and Travel Time
  - Employ geometric network files with speed limits, road direction, and turn restrictions
  - Search algorithm necessary to find shortest path on a road network while minimizing the travel cost (e.g. Dijkstra algorithm)

Distances from two locations

- 08609 (Hamilton, NJ) to 19104 (Philadelphia PA)
  - HAVERSINE 30.08 MILES
  - GEODIST 29.97 MILES
  - DRIVING DISTANCE 46 MILES
  - DRIVING TIME 49 MINUTES

Distances from two locations

- 02837 (Little Compton, RI) to 02840 (Newport, RI)
  - HAVERSINE 7.61 MILES
  - GEODIST 7.58 MILES
  - DRIVING DISTANCE 25 MILES
  - DRIVING TIME 42 MINUTES
Resources to calculate distances

• NAACCR Great Circle Distance Calculator
  – A SAS program that calculates the GCD distance between the locations of cases at the time of diagnosis and the locations of treatment facilities.
  – Designed to be used with the NAACCR v10 or v11 record layout file
  – Can use either source (unconsolidated) or consolidated case records as input.

Resources to calculate distances

• NAACCR/Komen road network travel distance/time application
  – Calculates shortest driving distance and travel time from origin to destination for batches of records for the entire US and Canada
  – Source data: NAVTEQ street shapefiles
  – Web service allows users to upload encrypted data containing origin and destination latitude and longitude coordinates
  – Outputs shortest travel distance and travel time

Resources to calculate distances

• Google Maps
  – Obtain shortest driving distance and travel time from origin to destination.
  – Possible to obtain distance/travel time based on public transportation and walking routes.
  – Programs can be written to send and receive batches of records for the entire US and Canada
    • SAS or Python
  – *Data not encrypted—best for public use data
    • Distance from census block group to closest hospital
Exercise 4: Calculating Distances Between Locations

Example: Census Tract Centroids To Hospitals

NAACCR Central Registry Webinar: GIS Applications

- Wrap-up
- Questions??

Next Month...

- Collecting Cancer Data: Esophagus and Stomach — June 3, 2010

...July Webinar

- Using CINA Data in Cancer Surveillance Activities — July 8, 2010
2010-2011 Webinar Series

- Available for purchase