Challenges and Solutions in Mapping Small Area Health Data

Thomas Talbot
2018 CSTE Pre-conference Workshop
West Palm Beach, FL  June 10, 2018

- Geocoding National & New York State Mortality Data
- Examples of displaying uncertainty on maps
- How to classify data (cut points)
- Selecting map colors
- Geographic Aggregation
Geocoding

• *Geocoding* is the process of transforming an address to a location on the Earth's surface.

• The location can either be a point (latitude, longitude) or an area (census tract)
US Small-Area Life Expectancy Estimates Project

Geocoded the National Death Certificate Data
2010-2015

• HUD, through its Geocode Service Center, validated and geocoded the death certificate data

• 96.1% of addresses were geocoded to the census tract based on the street address or 9-digit ZIP Code
% of 2010-2015 Records Geocoded*

* ME and WI: % of records that were geocoded 2011-2015

From a webinar, Geocoding the nation’s mortality data by Loraine Escobedo, NCHS 5/15/2018
Percent Non-geocoded hospital records by ZCTA

Rural areas are likely to have a greater percentage of ungeocoded records than non-rural areas.
Incomplete geocoding of mortality data can lead to bias estimates of life expectancy.

If we only geocode 50% of the deaths we add 10 years to the life expectancy estimate.

Assumes area has a life expectancy the same as the upstate NY area & both the deaths geocoded & deaths not geocoded have the same age distribution.
Positional error in the geocoded data may bias results. Estimating positional error is time consuming involving random sampling of addresses and interactive geocoding.

Figure 4
TIGER based positional error cumulative density distribution. This plot can be used to estimate the percent error at any distance for the three density classifications. The vertical dashed lines show the error distance at the 95th percentile.
Percent Mismatch

Percent of geocoded results that were not assigned to the same county based on address, as listed in the hospital discharge data county record.
New York State SCALE Project
Geocoded 99.97%
2008-2012 NY State deaths (excluding NYC) to the census tract*

Stepwise process

96.9% geocoded using either residential rooftop locations or street line files
0.6% geocoded using address linked to hospital files
0.2% geocoded if the ZIP code was completely contained within a tract
0.6% geocoded if house number was missing but the street was completely contained in tract
1.2% interactively geocoded
0.5% imputed to tract based on ZIP code, town or county, race, ethnicity and age


Choosing cut points, color & displaying uncertainty on maps

• **Who is the audience?**
  – Will they understand the map?
  – What will they use the map for?

• **What do you want to highlight?**
  – Changes in time
  – Compare data with other jurisdictions
  – Examine similarities in geographic patterns between multiple indicators
  – Highlight areas with highest rates
  – Show magnitude of the differences in rates

- Shows the highest and lowest rates in each time period.
- Shows changes in rates over time.
Different methods for classifying data for choropleth maps

These methods are often programmed into the GIS Software
Quantiles

Equal number of areas in each class

**Advantages:**

Quantiles can be useful at comparing relative rankings between indicators.
Rankings can also be looked at over time.

**Disadvantages:**

Because equal numbers of areas are placed in each class. Areas with similar rates can be
Placed in adjacent classes, or areas with widely different values can be put in the same class.

If each state uses quantile methods rates can not be
compared between states since the cut points will not be the same.

**Solutions:**

Increase number of classes
Life Expectancy by Albany County Regions 2008-2012

Example: Quantiles

Interval
5.8 years
2.1
1.3
1.6
3.7

University of Albany

T. Talbot, 2018
Equal Intervals

Min & max value used so that each class has the same interval

Advantages:

Useful in comparing patterns of absolute values and locations of outliers

Disadvantages:

Maybe difficult to compare across indicators since there are not equal numbers in each class
The class intervals can be very large with most the data grouped in the same class.

Solutions

Use equal intervals for all the classes in the middle of the distribution but for the tails use a different interval
Natural Breaks (Jenks’ Method)
Variance is minimized within classes and maximized among classes

**Advantages:**
Maximizes differences between classes

**Disadvantages:**
Difficult to explain to public
Data-specific classifications, not useful for comparing multiple maps built from different underlying information.
The class intervals can be very large with most of the data grouped in the same class.

**Solutions**
Use method to better understand your data if you are manually deciding on class breaks.
Standard Deviation
Class breaks are determined by standard deviations from the mean

**Advantages:**

Used to help identify areas that are above and below the mean.
Can be used to compare multiple indicators.
Typically will have more areas in the class around the mean than at either tail of the distribution.
Good for identifying extreme values

**Disadvantages:**

Different number of areas in each class
Comparing data across states or across time difficult since the class breaks will change with the data.

**Solutions**

Determine class breaks based on nationwide data & across all time periods.
Percent Difference from the Mean
Class breaks are determined by percent difference from the mean

**Advantages:**
Used to help identify areas that are above and below the mean.
Can be used to compare multiple indicators.
Easy to explain to the public.
Good for identifying extreme values.

**Disadvantages:**
Different number of areas in each class
Comparing data across states or across time difficult since the class breaks will change with the data.

**Solutions**
Determine class breaks based on nationwide data & across all time periods.
Example: Percent difference from the mean
ColorBrewer
A web tool for selecting colors for maps

The Team

Cynthia Brewer: Concept / Colors / Editor
Mark Harrower: Design / User Interface
Ben Sheesley: Design / User Interface
Andy Woodruff: Developer
David Heyman: Developer

http://colorbrewer2.org
Number of data classes: 9
Nature of your data:
- sequential
- diverging
- qualitative
Pick a color scheme:
- Multi-hue
- Single hue

Only show:
- colorblind safe
- print friendly
- photocopy safe

Context:
- roads
- cities
- borders

Background:
- solid color
- terrain

Color Transparency:
- color transparency

247,252,240
224,243,219
204,235,197
168,221,181
123,204,196
78,179,211
43,140,190
8,104,172
8,64,129
Displaying Uncertainty
Maps linked to graphs showing confidence intervals

https://gis.cancer.gov/tools/micromaps/
Geographic Aggregation

• Rates will be unstable in areas with small numbers of people & health events.

• One solution is to aggregate several years of data.

• Even after temporal aggregation we might not have enough data to produce reliable rates in some areas.

• Solution: aggregate small areas into larger areas until we have adequate data for reliable indicators
Selecting the geographic scale for calculating life expectancy in NY
5 years of death data 2008-2012

Geographic Aggregation

• Aggregated census tracts together so each region contained at least 60 deaths and had a standard error <2 years.

• Set our Geographic Aggregation Tool so no region crossed larger town or city boundaries.

• In rural areas regions were created to not cross county boundaries
Aggregation Tool

Original Block Data †

<table>
<thead>
<tr>
<th>Block</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>122300/2004</td>
<td>2</td>
</tr>
<tr>
<td>122300/2005</td>
<td>11</td>
</tr>
<tr>
<td>014500/3005</td>
<td>2</td>
</tr>
<tr>
<td>014500/3007</td>
<td>3</td>
</tr>
<tr>
<td>014500/3008</td>
<td>8</td>
</tr>
<tr>
<td>014500/3009</td>
<td>3</td>
</tr>
<tr>
<td>014500/3010</td>
<td>4</td>
</tr>
<tr>
<td>103202/2001</td>
<td>9</td>
</tr>
<tr>
<td>103202/2002</td>
<td>5</td>
</tr>
</tbody>
</table>

Regions

<table>
<thead>
<tr>
<th>Block</th>
<th>Cases</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>122300/2004</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>122300/2005</td>
<td>11</td>
<td>A</td>
</tr>
<tr>
<td>014500/3005</td>
<td>2</td>
<td>B</td>
</tr>
<tr>
<td>014500/3007</td>
<td>3</td>
<td>B</td>
</tr>
<tr>
<td>014500/3008</td>
<td>8</td>
<td>B</td>
</tr>
<tr>
<td>014500/3009</td>
<td>3</td>
<td>B</td>
</tr>
<tr>
<td>014500/3010</td>
<td>4</td>
<td>B</td>
</tr>
<tr>
<td>103202/2001</td>
<td>9</td>
<td>C</td>
</tr>
<tr>
<td>103202/2002</td>
<td>5</td>
<td>C</td>
</tr>
</tbody>
</table>
For further information on methods used to calculate life expectancy in New York State see:


Contact: Thomas Talbot, thmstalbot@gmail.com