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 Nations Death Certificate Data
2010-2015

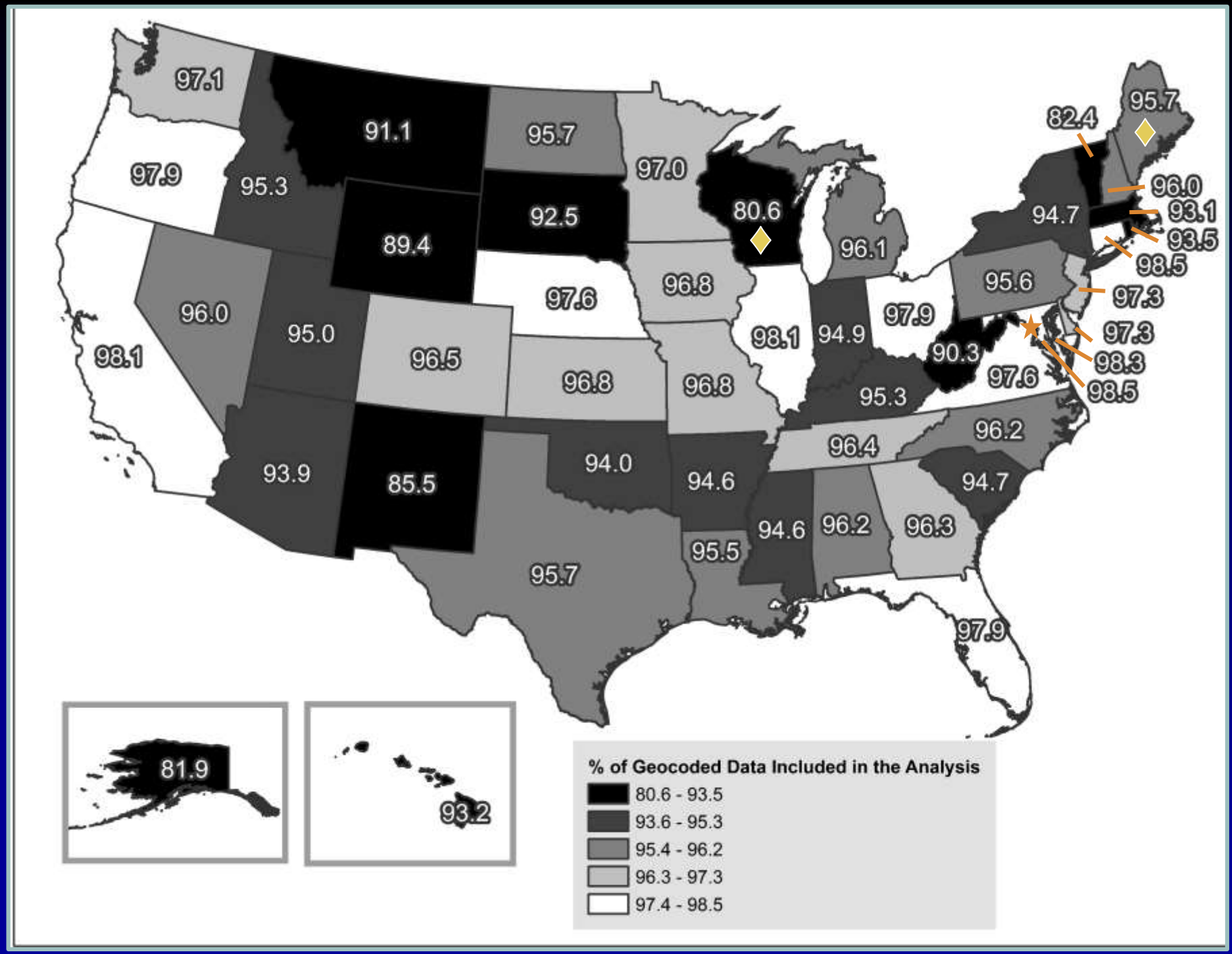
- HUD, through its Geocode Service Center, validated and geocoded the death certificate data

- 96.1% of addresses were geocoded to the census tract based on the street address or 9-digit ZIP Code

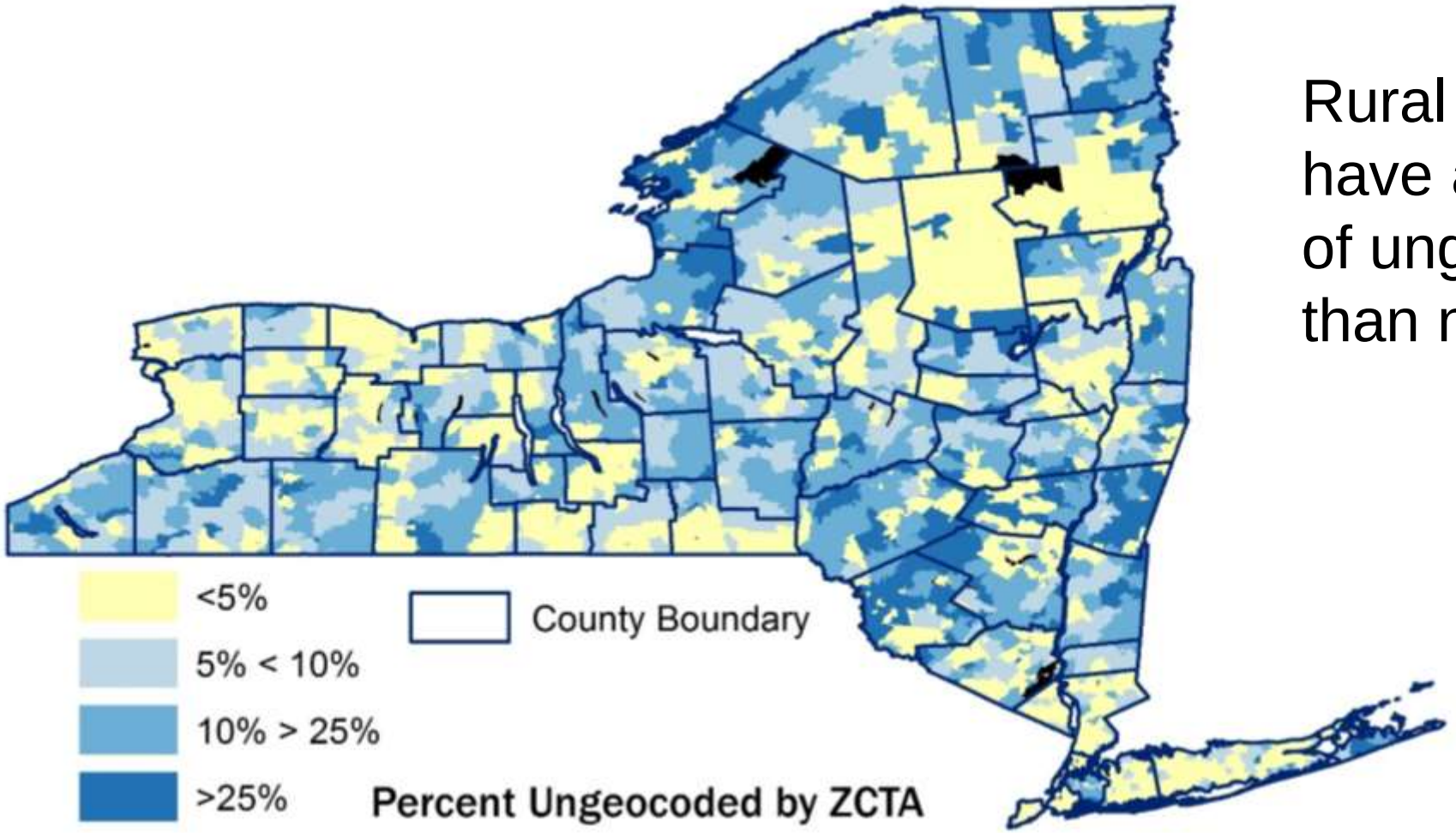
*% of 2010-2015 Records Geocoded**

* ***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



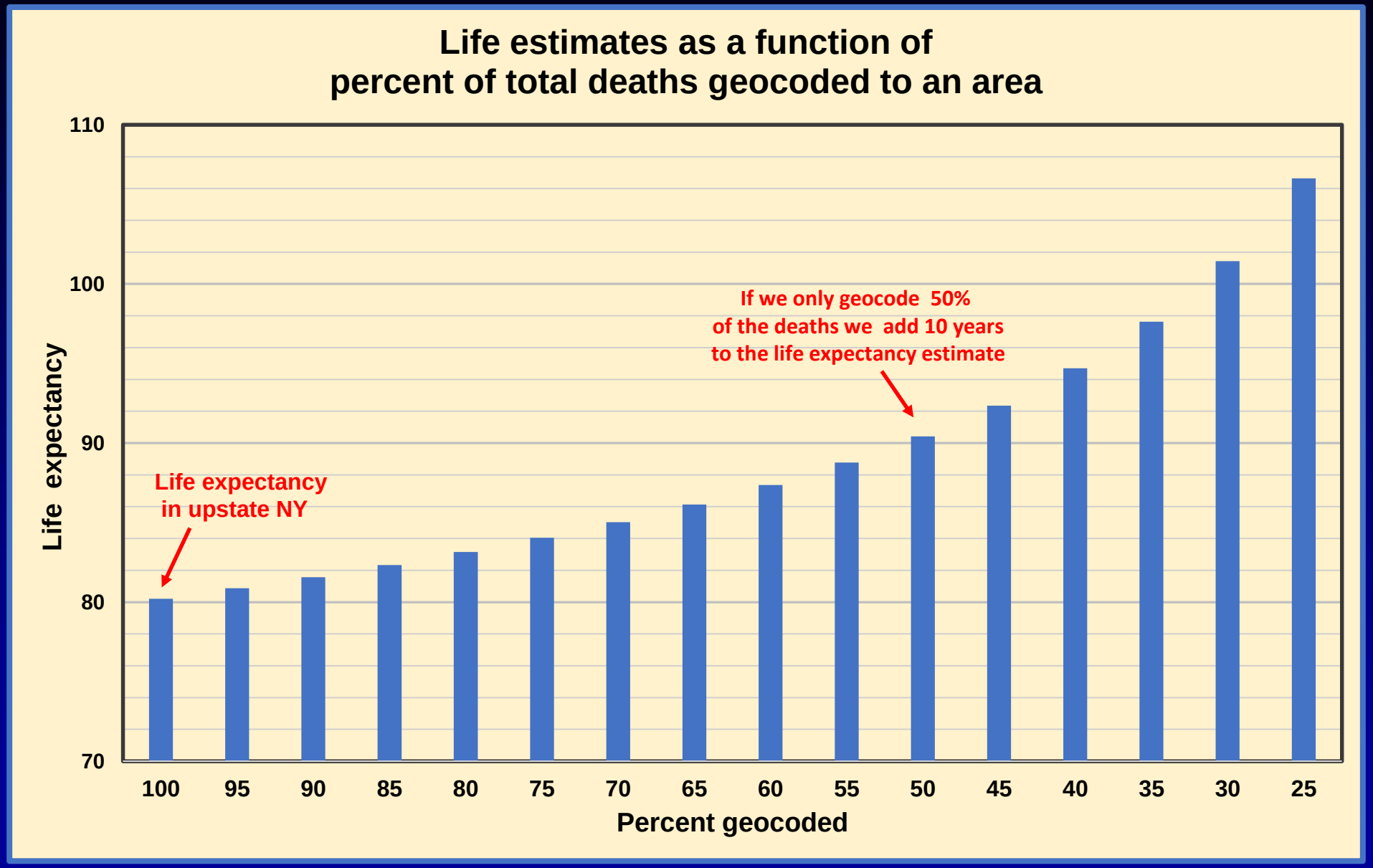
Percent Non-geocoded hospital records by ZCTA



Rural areas are likely to have a greater percentage of ungeocoded records than non-rural areas

NEW YORK STATE OF OPPORTUNITY. Department of Health

Incomplete geocoding of mortality data can lead to bias estimates of life expectancy



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.

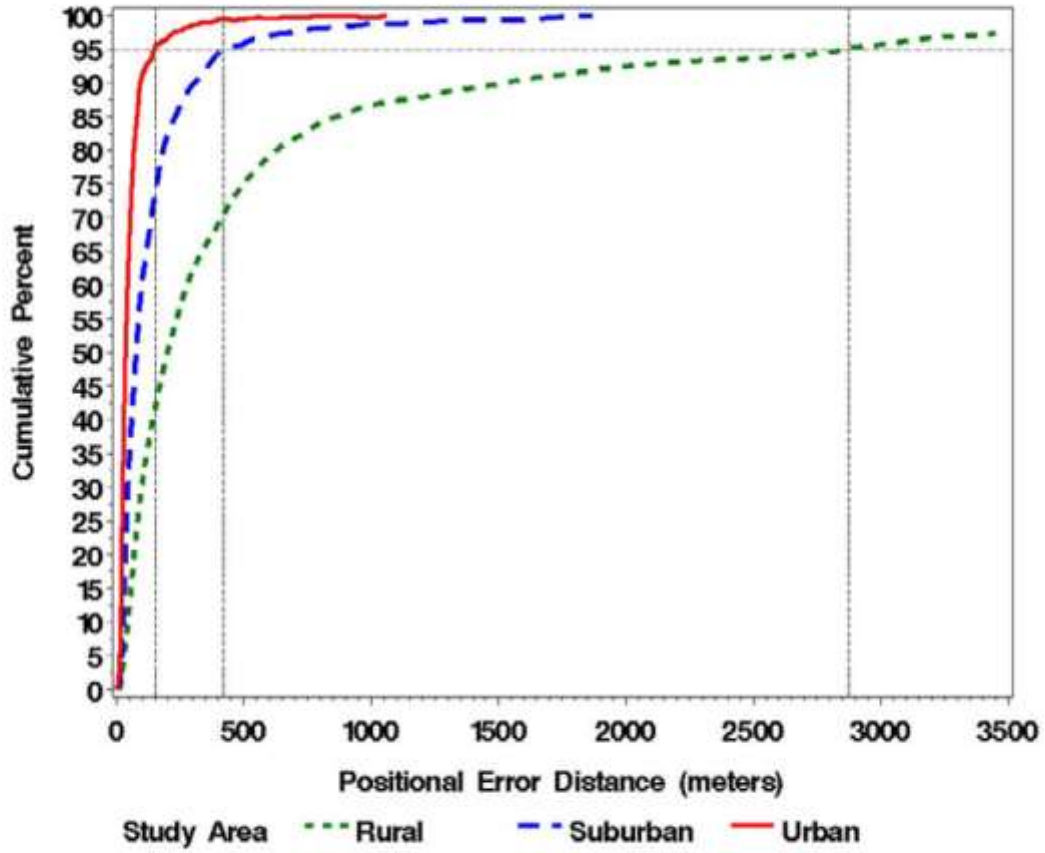
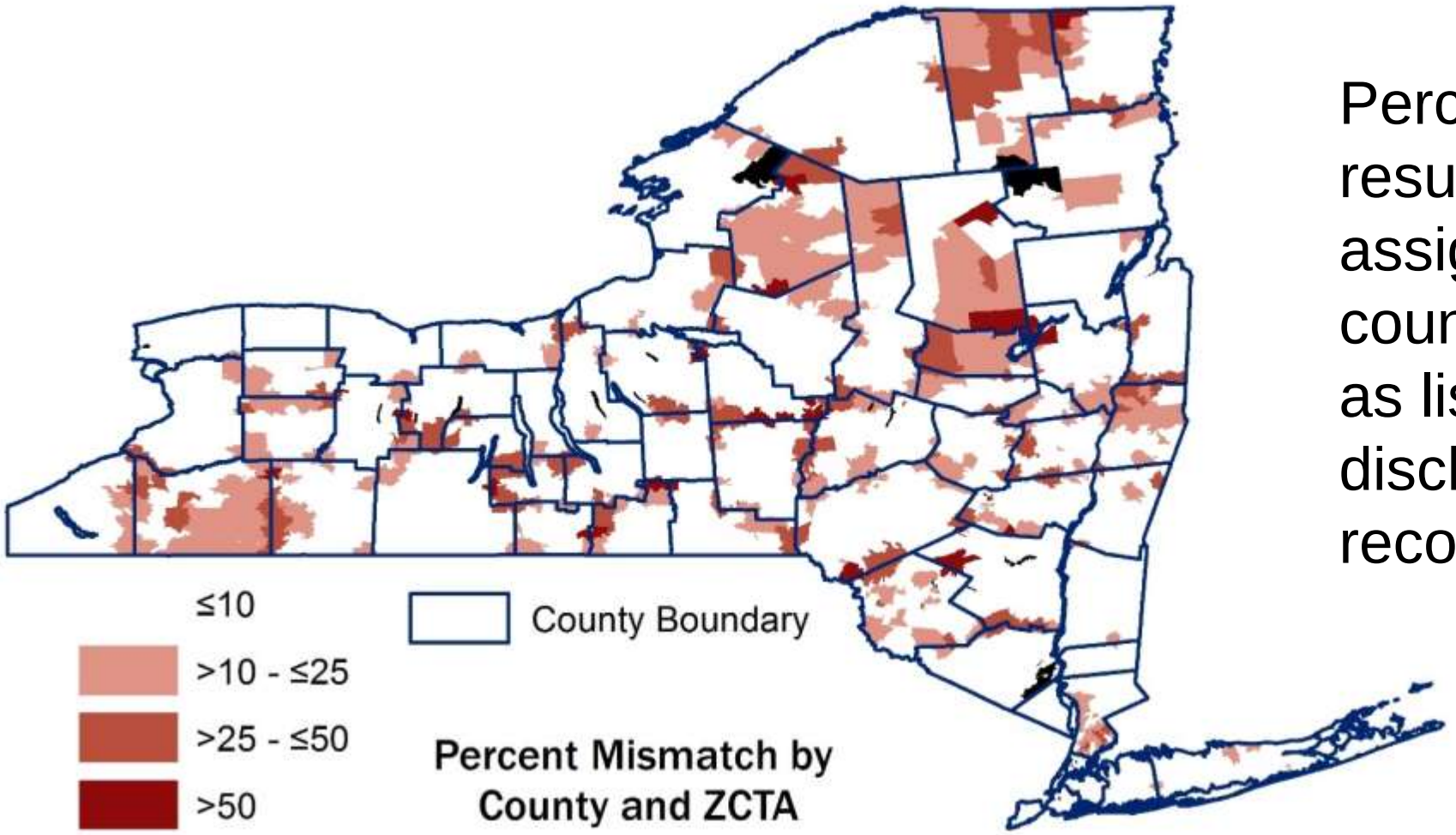


Figure 4
TIGER based positional error cumulative density distribution. This plot can be used to estimate the percent error at any distance for the three density classifications. The vertical dashed lines show the error distance at the 95th percentile.

Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2003, 2:10

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 OF OPPORTUNITY. Department of Health

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[1]

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[2]

**** Talbot et al. Calculating census tract-based life expectancy in New York state: a generalizable approach. Pop Health Metrics, 2018 Jan 26; 16(1):1***

1) NYS GIS Program Office Geocoding Services. New York State GIS Program Office. 2016. https://gis.ny.gov/gisdata/inventories/details.cfm?DSID=1278. Accessed 19 Dec 2016.

2) Henry KA, Boscoe FP. Estimating the accuracy of geographical imputation. Int J Health Geogr. 2008; https://doi.org/10.1186/1476-072X-7-3.

Choosing cut points, color & displaying uncertainty on maps

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

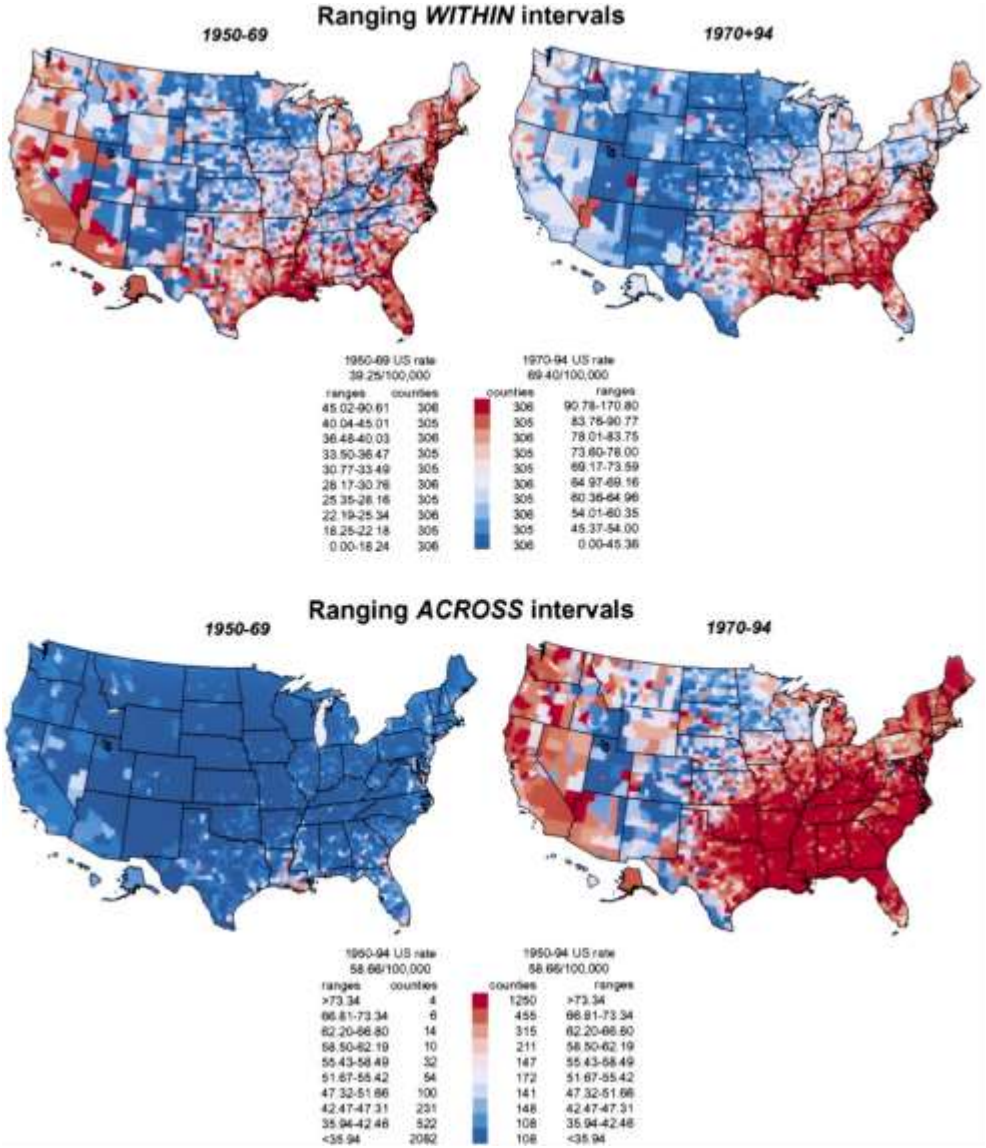
- **What do you want to highlight?**
  - Changes in time
  - Compare data with other jurisdictions
  - Examine similarities in geographic patterns between multiple indicators
  - Highlight areas with highest rates
  - Show magnitude of the differences in rates

Showing data for different time periods

Shows the highest and lowest rates in each time period

Shows changes in rates over time

Lung cancer mortality rates by county for the two time intervals, from 1950 through 1969 and from 1970 through 1994



Grauman DJ, Tarone RE, Devesa SS, Fraumeni JF.Alternate Ranging Methods for Cancer Mortality Maps, J Natl Cancer Inst. 2000;92(7):534-543. doi:10.1093/jnci/92.7.534 J Natl Cancer Inst

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:

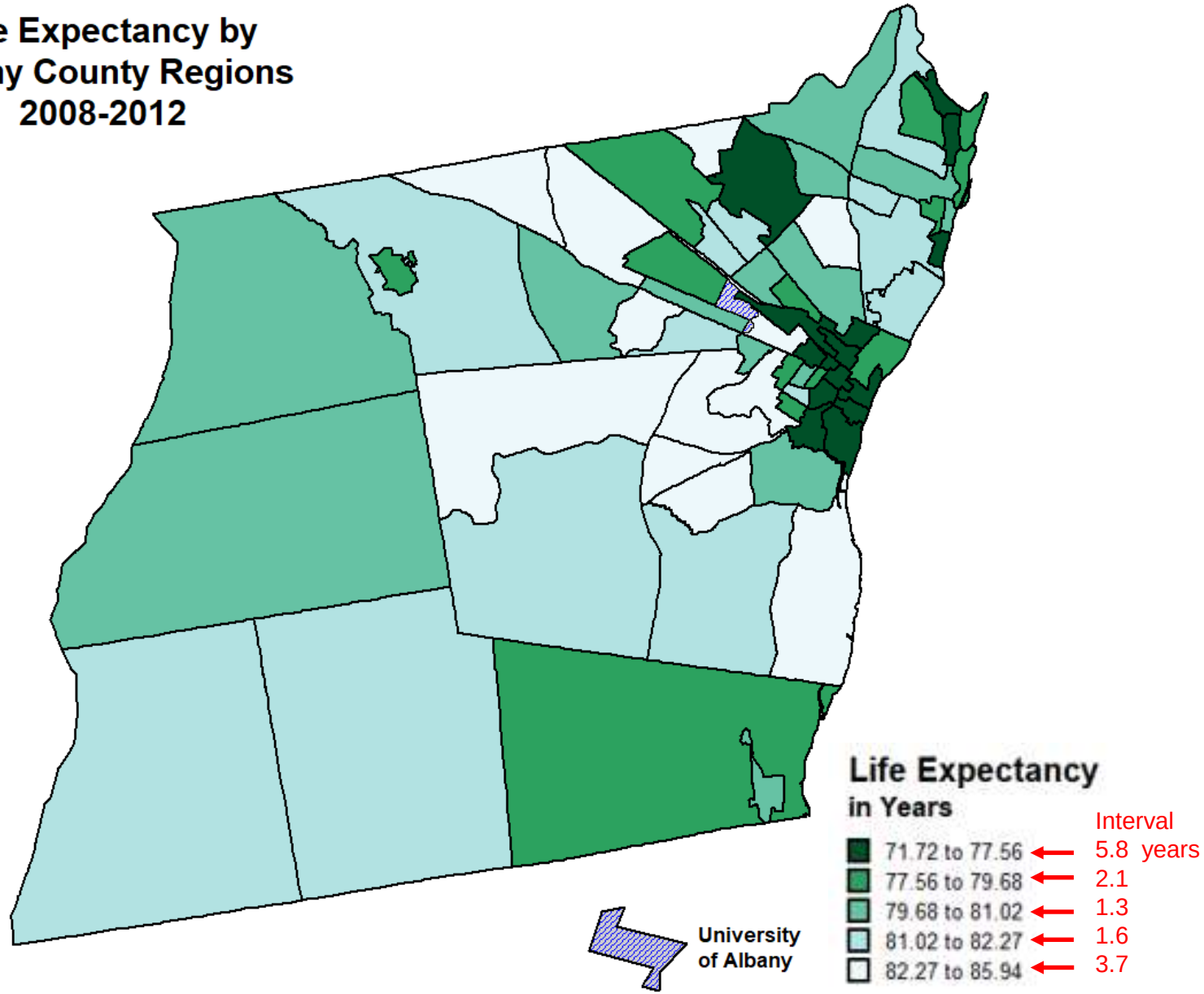
Because equal numbers of areas are placed in each class. Areas with similar rates can be Placed in adjacent classes, or areas with widely different values can be put in the same class.

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



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

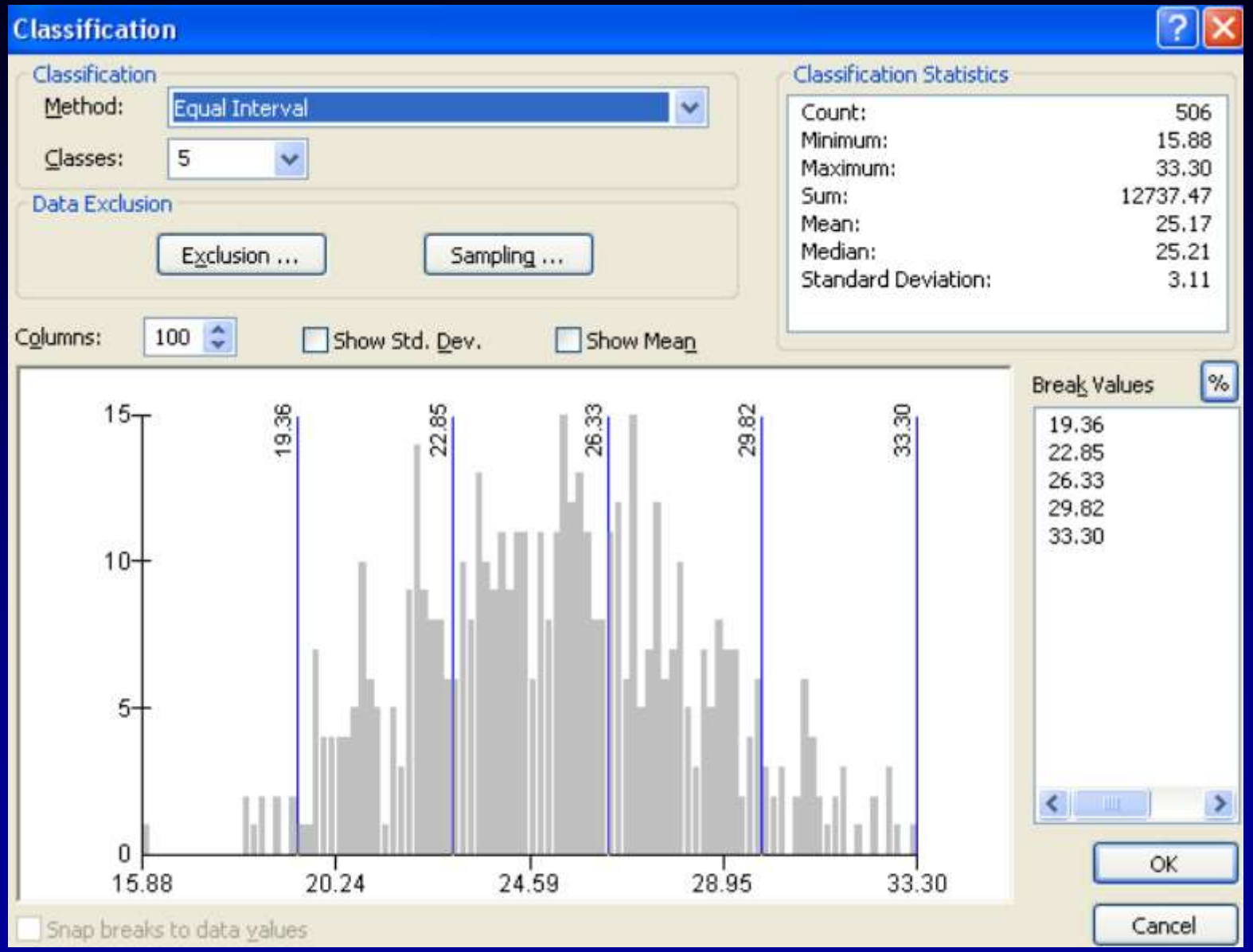
Disadvantages:

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


Classification
Classification
Method:
Equal Interval
Classes:
5
Data Exclusion
Exclusion ...
Sampling ...
Classification Statistics
Count: 506
Minimum: 15.88
Maximum: 33.30
Sum: 12737.47
Mean: 25.17
Median: 25.21
Standard Deviation: 3.11
Columns:
100
Show Std. Dev.
Show Mean
15
10
5
0
19.36
22.85
26.33
29.82
33.30
15.88
20.24
24.59
28.95
33.30
Break Values
%
19.36
22.85
26.33
29.82
33.30
OK
Cancel
Snap breaks to data values

Natural Breaks (Jenks' Method)

Variance is minimized within classes and maximized among classes

Advantages:

Maximizes differences between classes

Disadvantages:

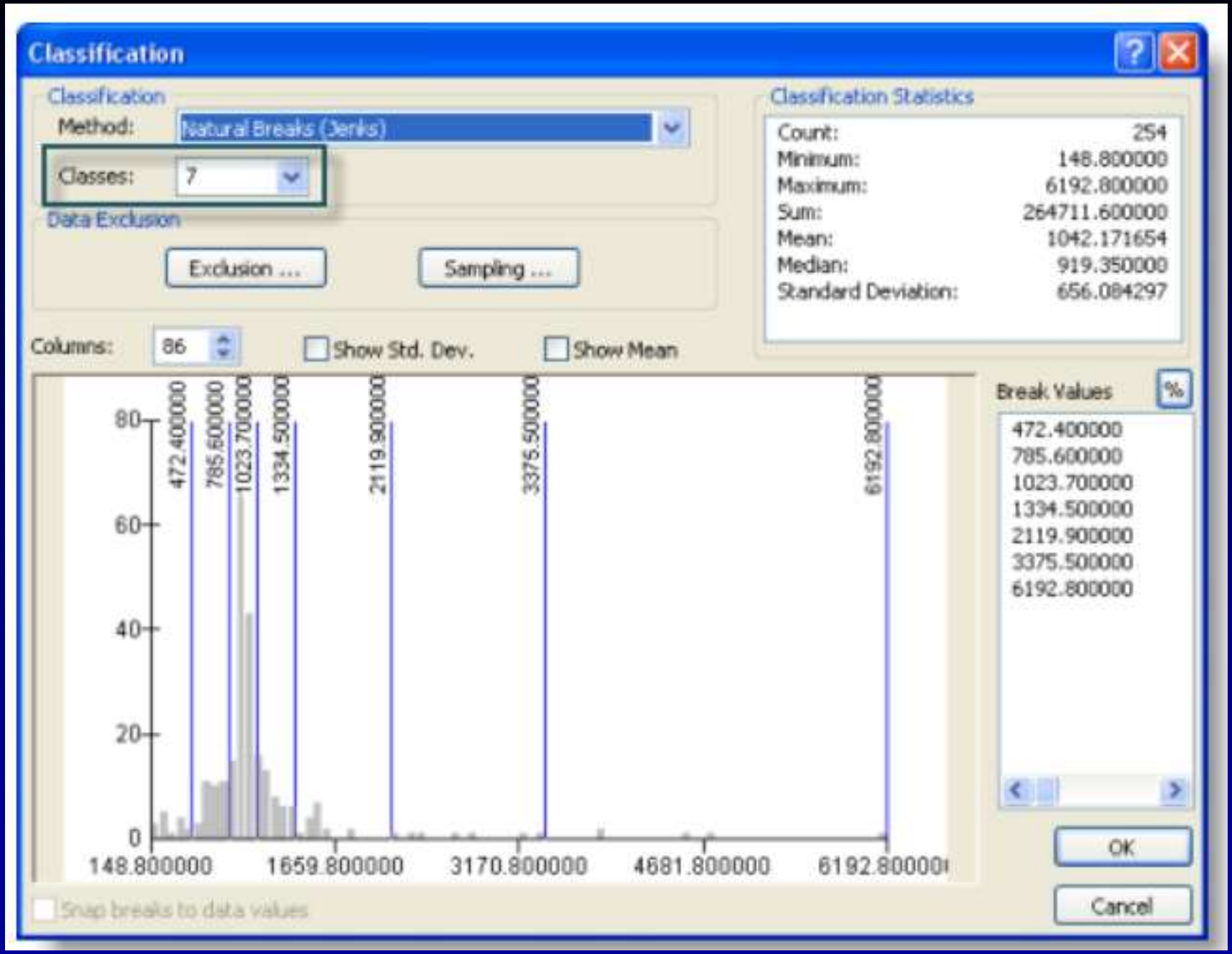
Difficult to explain to public

Data-specific classifications, not useful for comparing multiple maps built from different underlying information.

The class intervals can be very large with most the data grouped in the same class.

Solutions

Use method to better understand your data if you are manually deciding on class breaks.


Classification
Classification
Method:
Natural Breaks (Jenks)
Classes:
7
Data Exclusion
Exclusion ...
Sampling ...
Classification Statistics
Count: 254
Minimum: 148.800000
Maximum: 6192.800000
Sum: 264711.600000
Mean: 1042.171654
Median: 919.350000
Standard Deviation: 656.084297
Columns:
86
Show Std. Dev.
Show Mean
472.400000
785.600000
1023.700000
1334.500000
2119.900000
3375.500000
6192.800000
80
60
40
20
0
148.800000
1659.800000
3170.800000
4681.800000
6192.800000
Break Values
%
472.400000
785.600000
1023.700000
1334.500000
2119.900000
3375.500000
6192.800000
OK
Cancel
Snap breaks to data values

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:

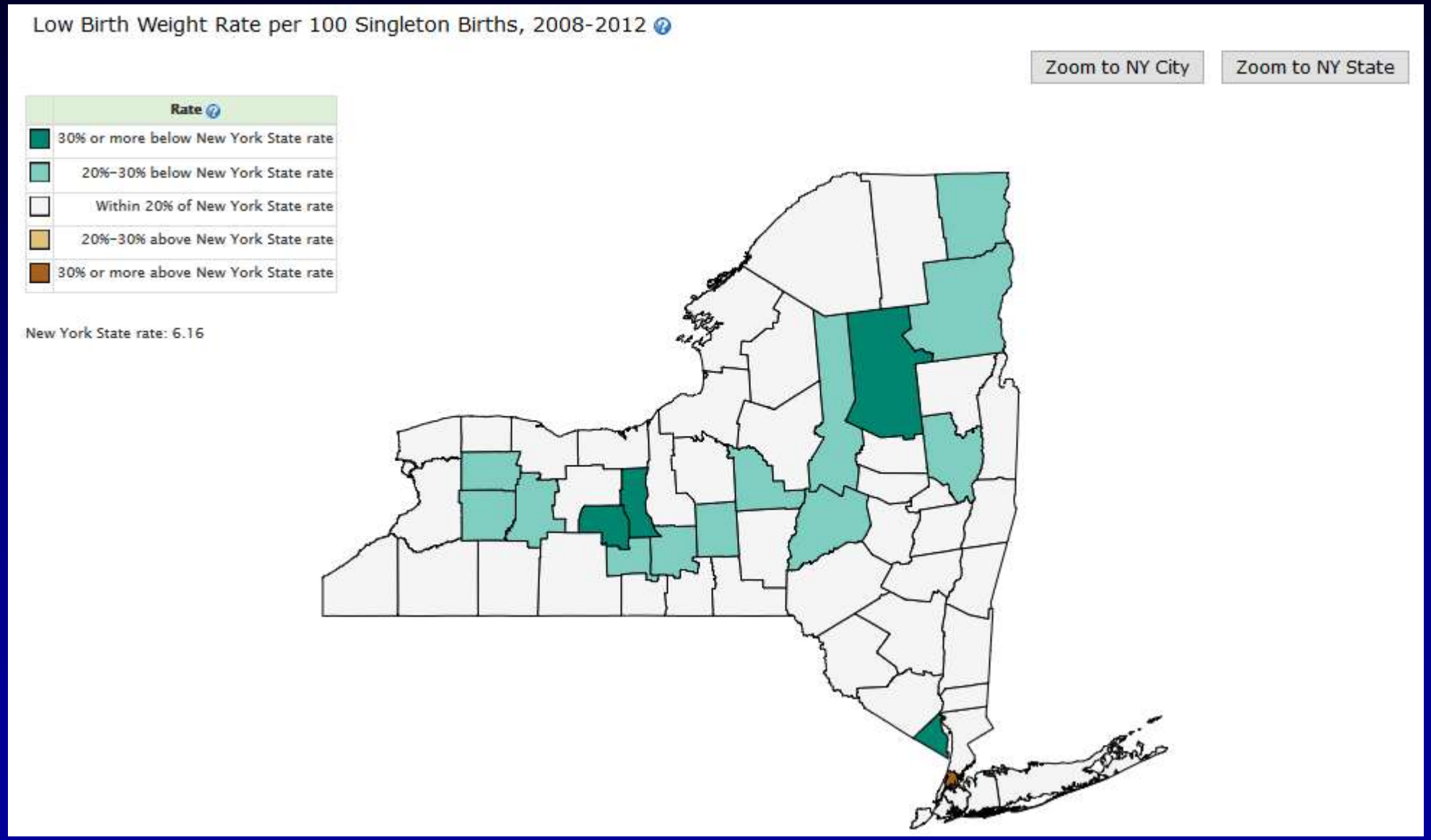
Different number of areas in each class

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

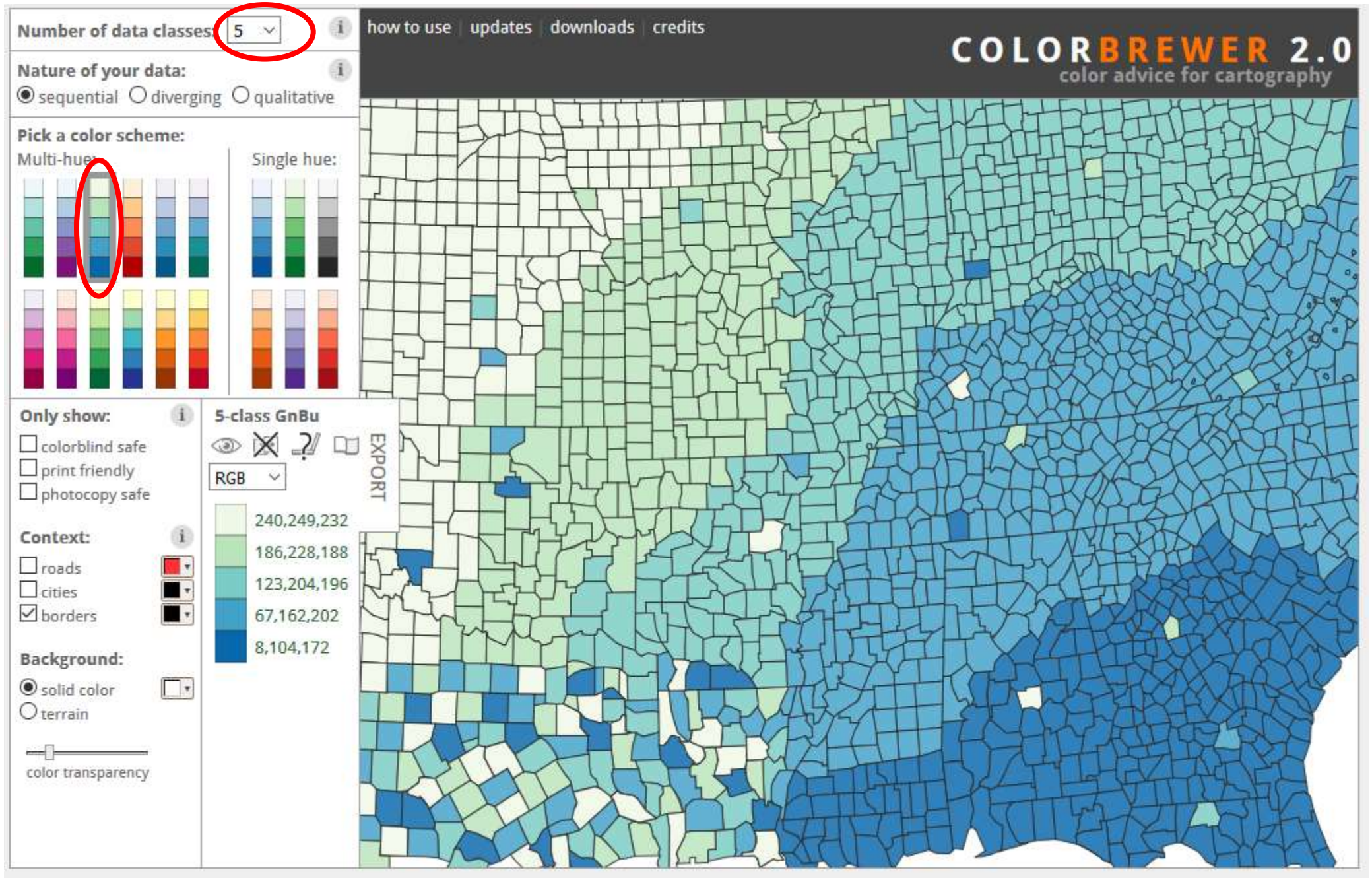
Mark Harrower: Design / User Interface

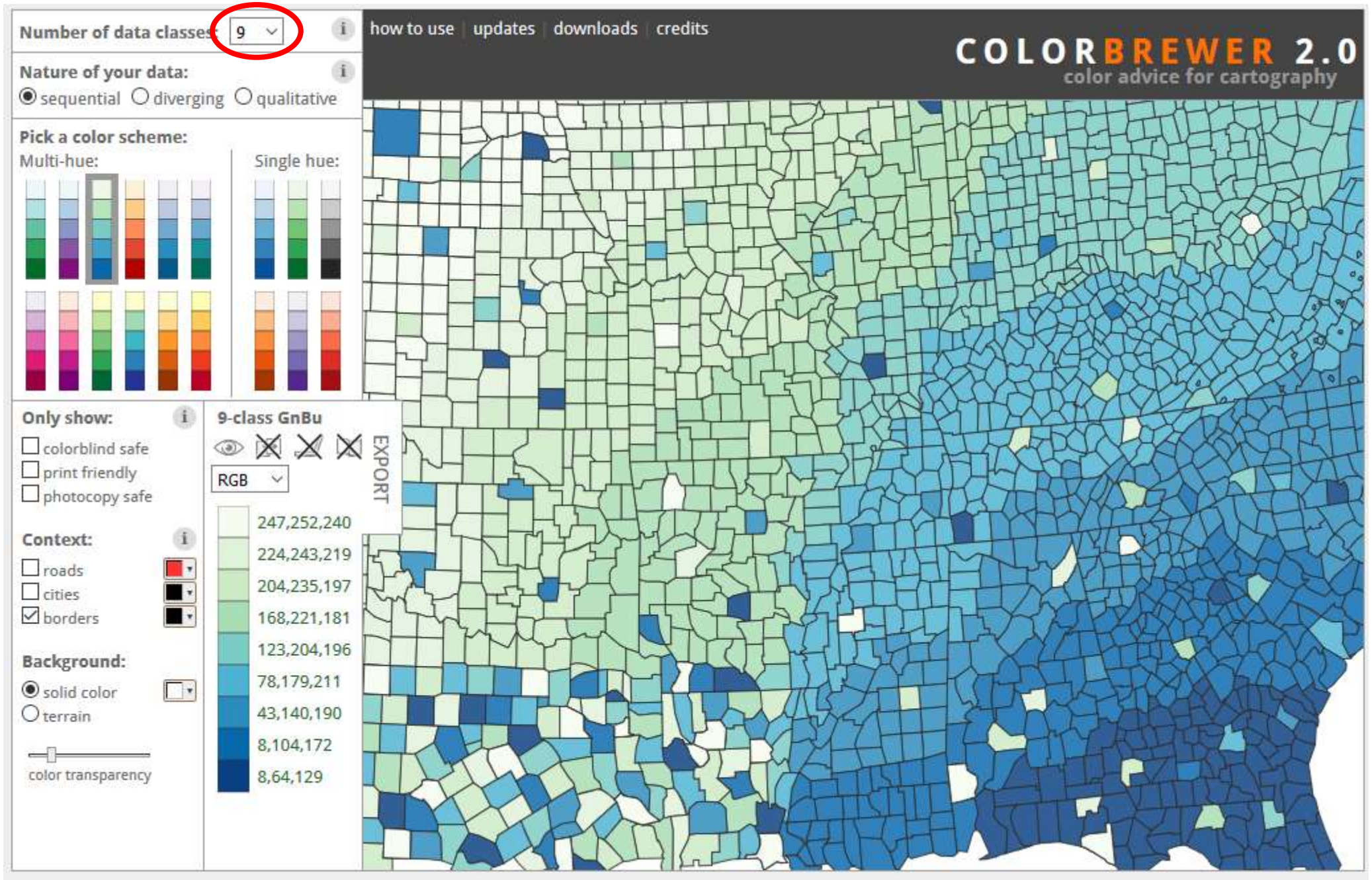
Ben Sheesley: Design / User Interface

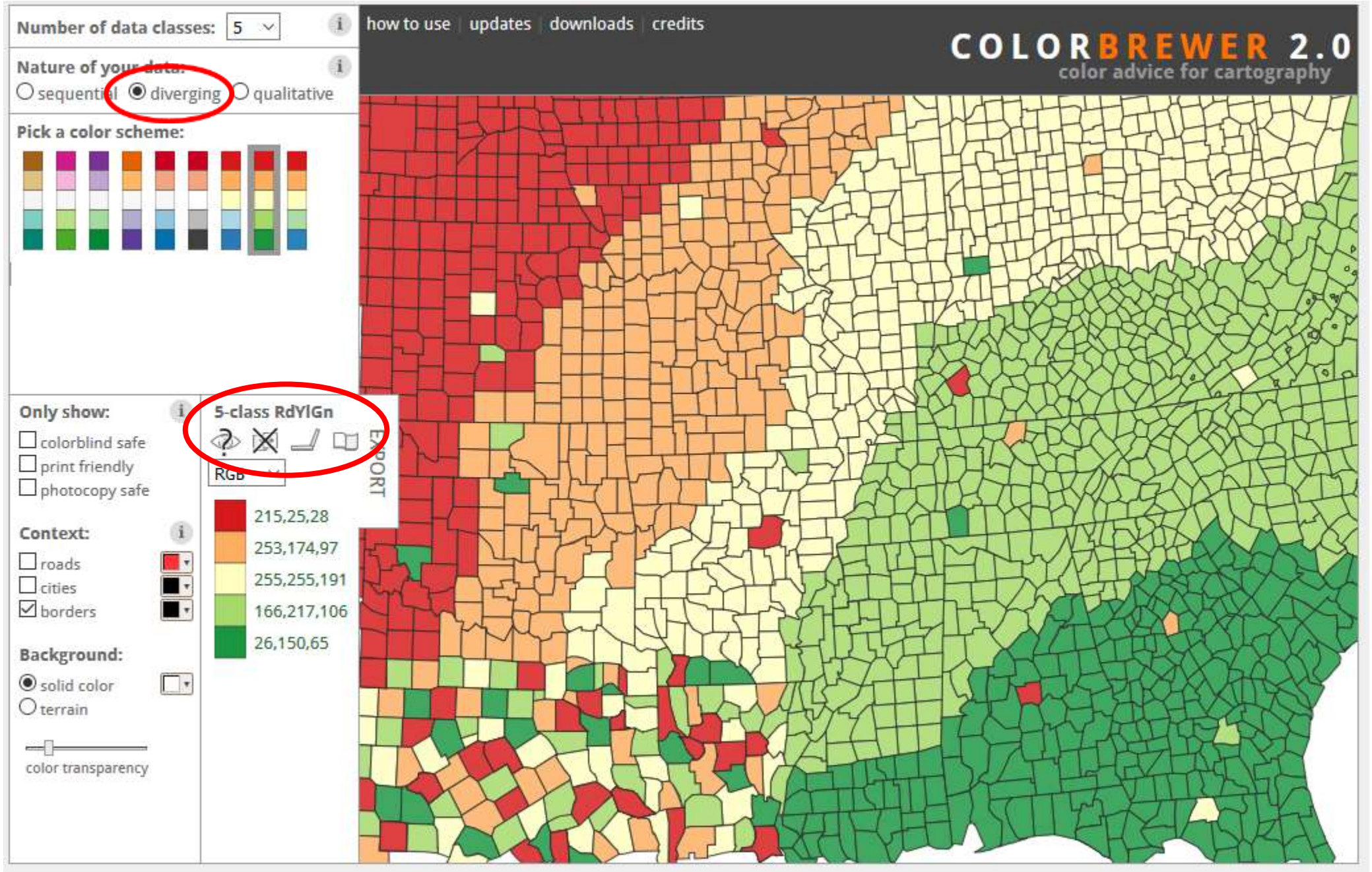
Andy Woodruff: Developer

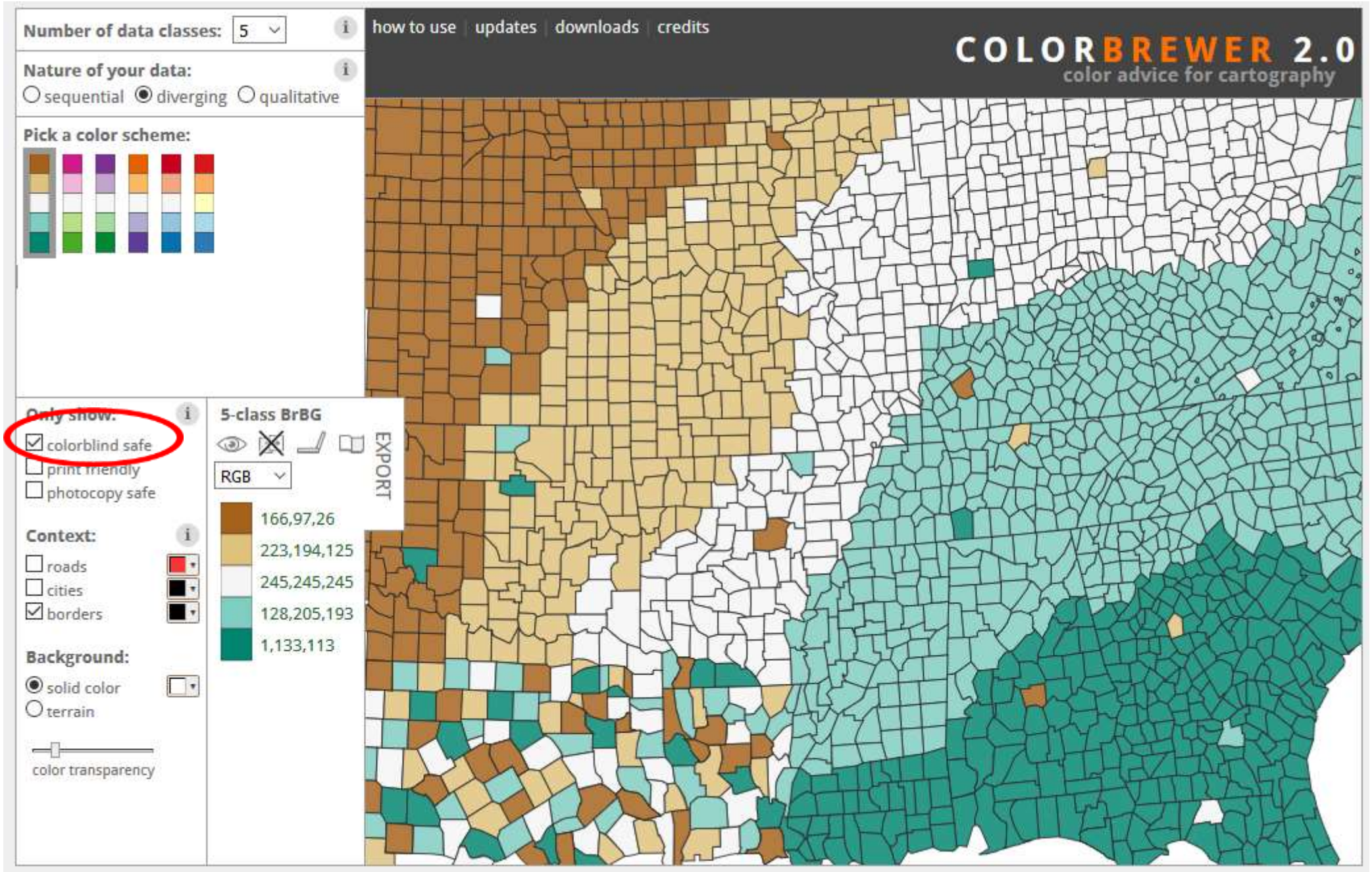
David Heyman: Developer

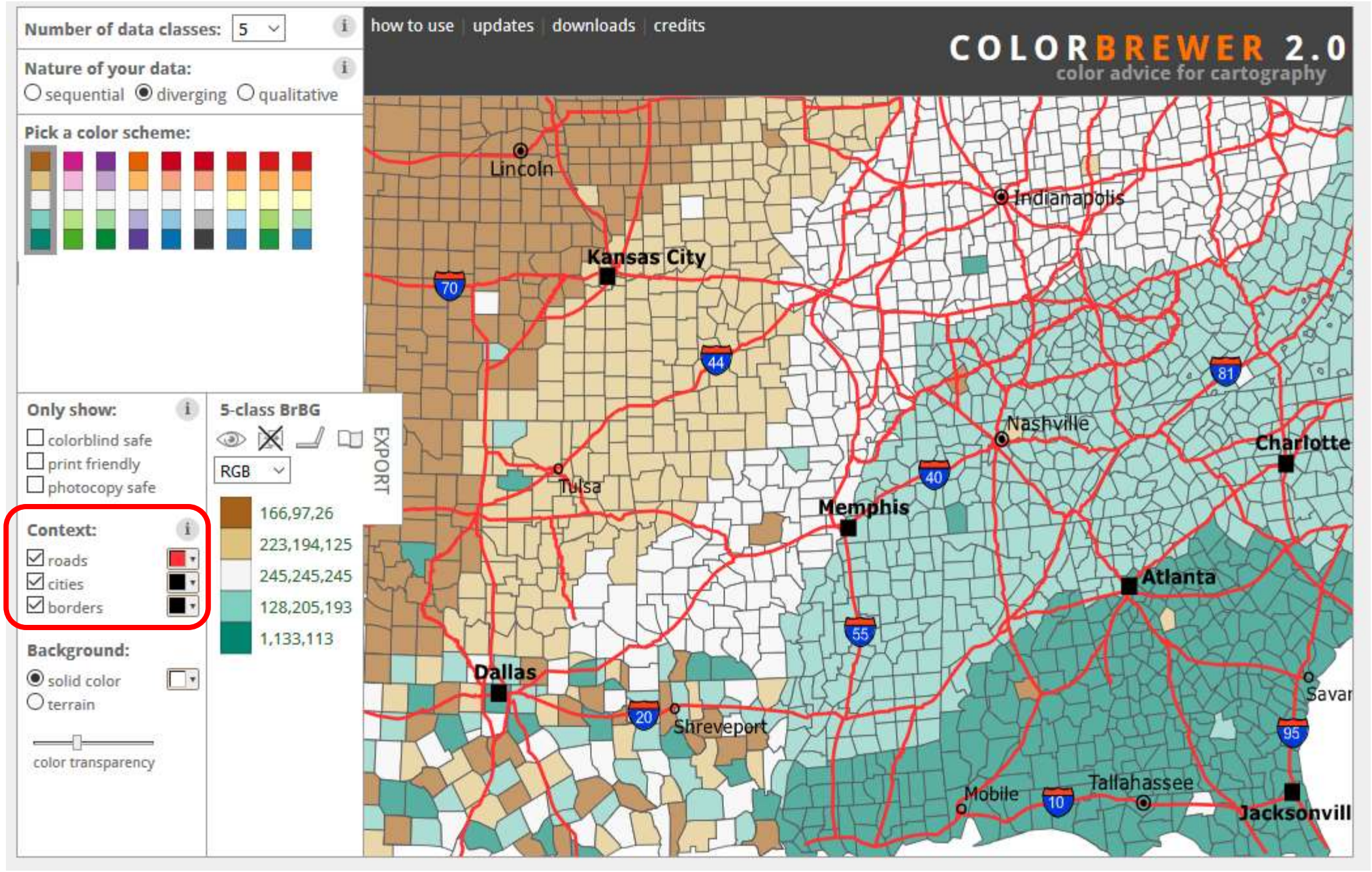
http://colorbrewer2.org

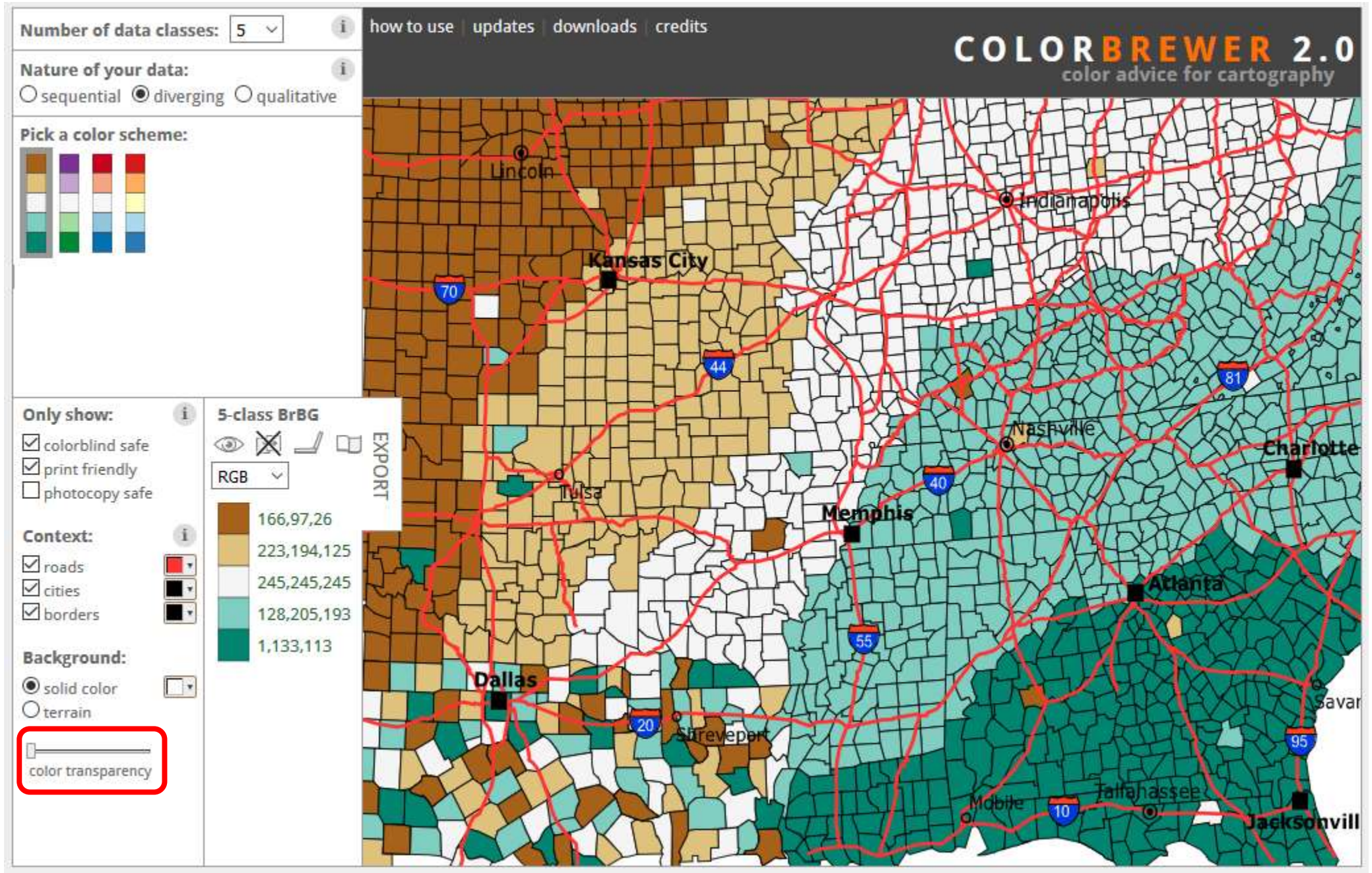

Number of data classes: 5
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
5-class GnBu
RGB
EXPORT
240,249,232
186,228,188
123,204,196
67,162,202
8,104,172
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography


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
9-class GnBu
RGB
EXPORT
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
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography

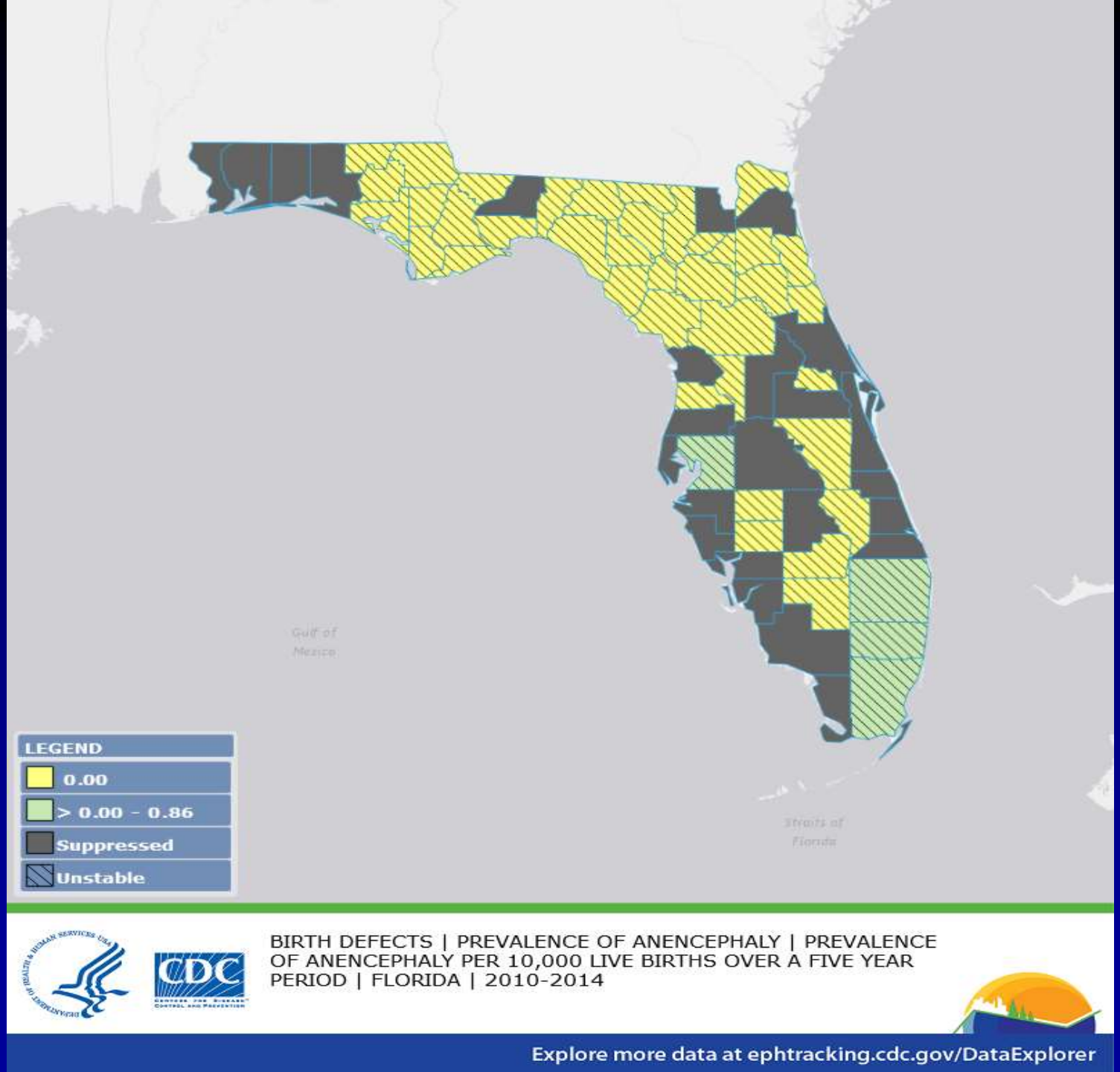

Number of data classes: 5
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
5-class RdYlGn
RGB
EXPORT
215,25,28
253,174,97
255,255,191
166,217,106
26,150,65
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography


Number of data classes: 5
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
5-class BrBG
RGB
EXPORT
166,97,26
223,194,125
245,245,245
128,205,193
1,133,113
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography

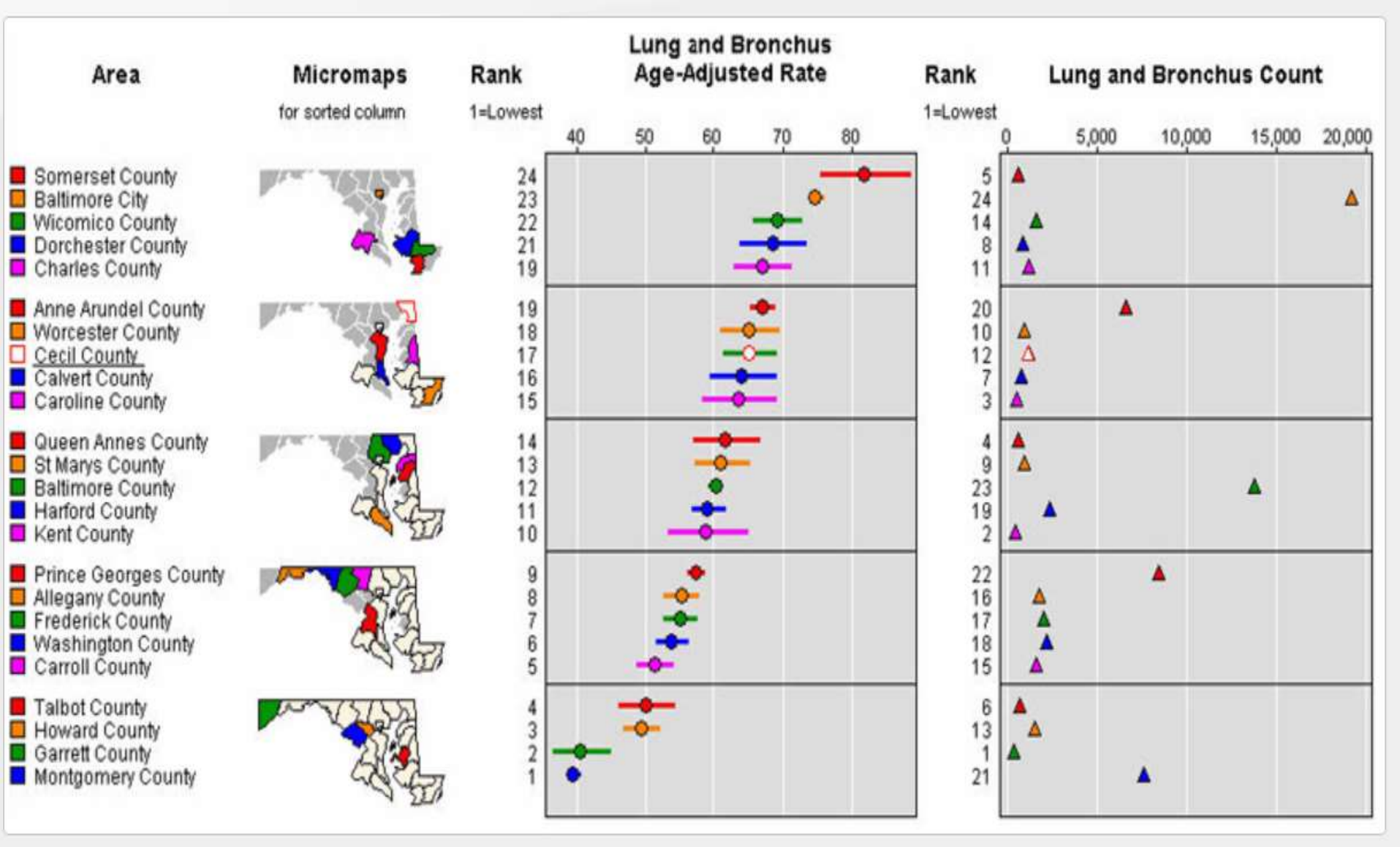

Number of data classes: 5
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
5-class BrBG
RGB
EXPORT
166,97,26
223,194,125
245,245,245
128,205,193
1,133,113
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography
Lincoln
Kansas City
Indianapolis
Nashville
Charlotte
Tulsa
Memphis
Atlanta
Dallas
Shreveport
Mobile
Tallahassee
Jacksonvill
70
44
40
81
55
20
95
10


Number of data classes: 5
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
5-class BrBG
EXPORT
RGB
166,97,26
223,194,125
245,245,245
128,205,193
1,133,113
Lincoln
Kansas City
Indianapolis
Nashville
Charlotte
Tulsa
Memphis
Atlanta
Dallas
Shreveport
Mobile
Tallahassee
Jacksonvill
70
44
81
40
55
20
10
95

Displaying Uncertainty



Maps linked to graphs showing confidence intervals



National Cancer Institute

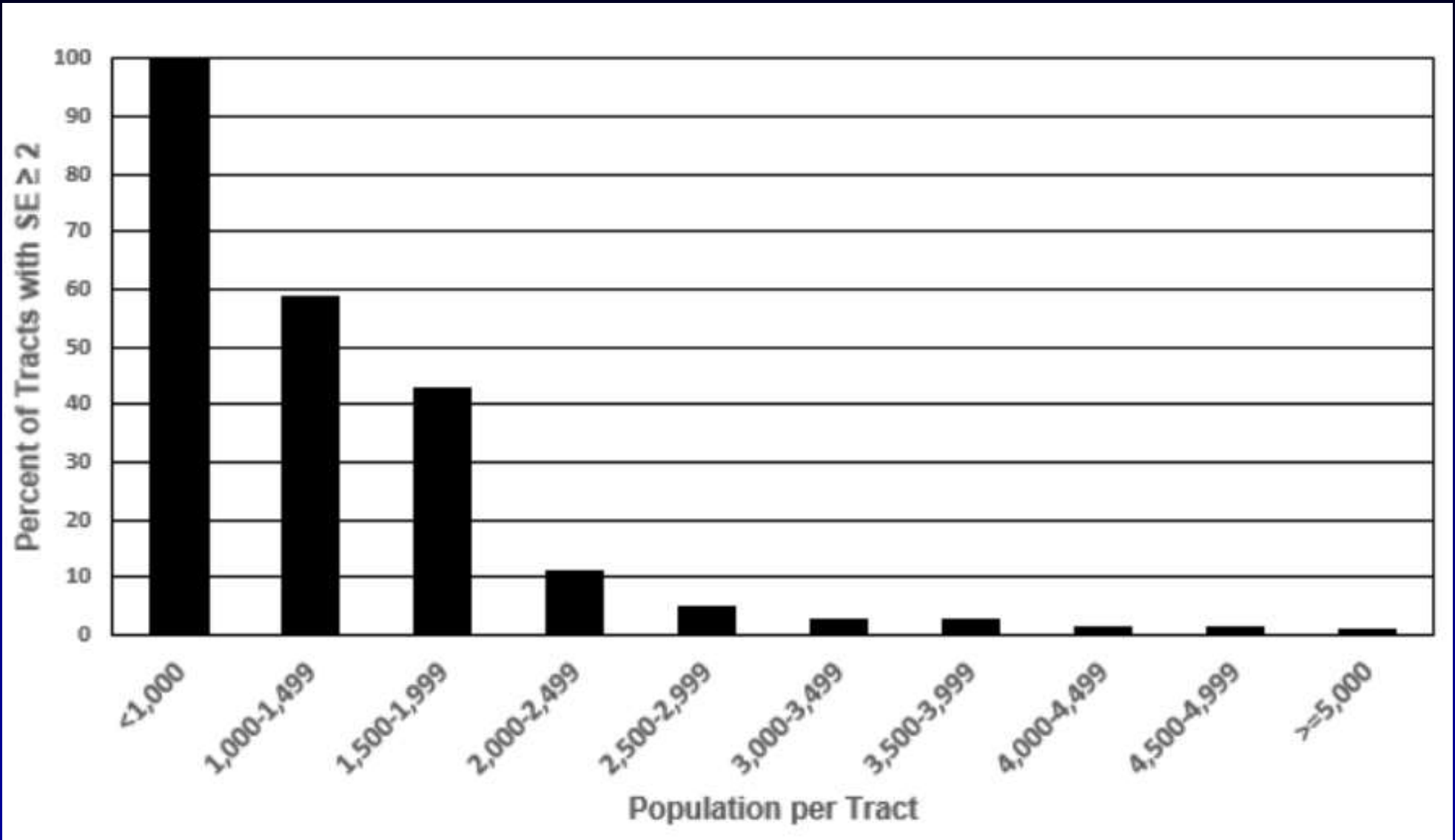
https://gis.cancer.gov/tools/micromaps/

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

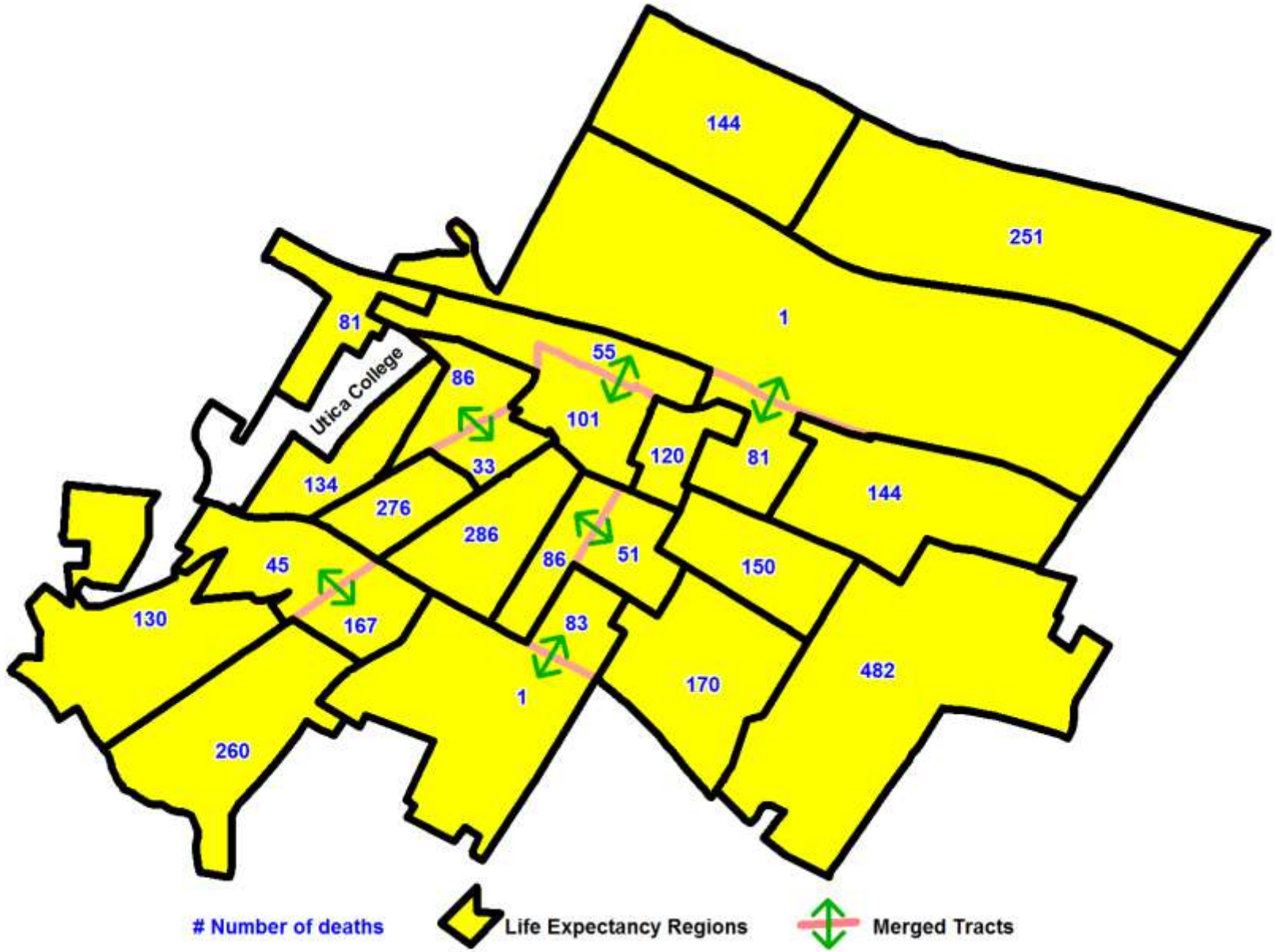
5 years of death data 2008-2012



Talbot et al. Calculating census tract-based life expectancy in New York state: a generalizable approach. Pop Health Metrics, 2018 Jan 26; 16(1):1

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



Talbot et al. Calculating census tract-based life expectancy in New York state: a generalizable approach. Pop Health Metrics, 2018 Jan 26; 16(1):1

Aggregation Tool

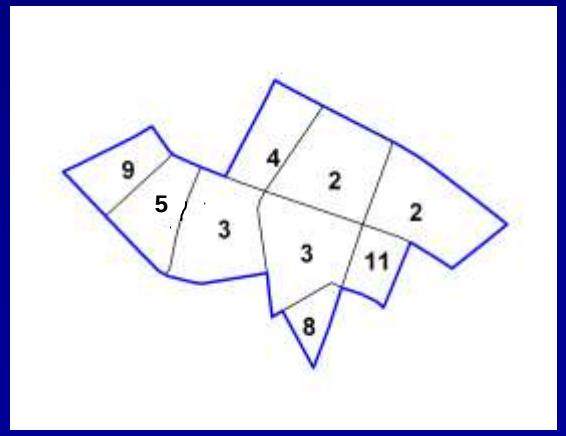
Original Block Data †

Block	Cases
122300/2004	2
122300/2005	11
014500/3005	2
014500/3007	3
014500/3008	8
014500/3009	3
014500/3010	4
103202/2001	9
103202/2002	5

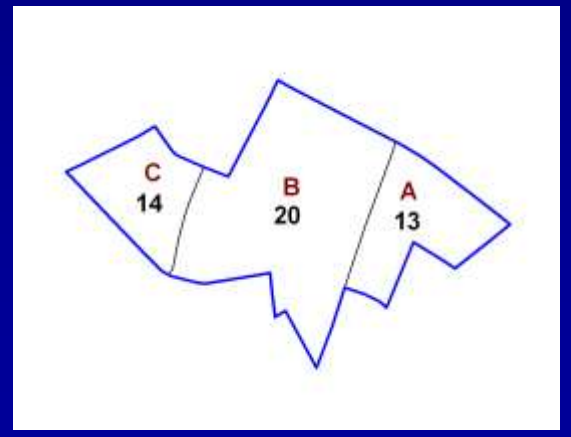
SAS or R Tool

Regions

Block	Cases	Region
122300/2004	2	A
122300/2005	11	A
014500/3005	2	B
014500/3007	3	B
014500/3008	8	B
014500/3009	3	B
014500/3010	4	B
103202/2001	9	C
103202/2002	5	C



Cases	Region
13	A
20	B
14	C



www.albany.edu/faculty/ttalbot/GAT/

Developed by Thomas O. Talbot and Gwen D. LaSelva.
Geographic Aggregation Tool, Version 1.2, New York State Health Department, Albany NY

For further information on methods used to calculate life expectancy in New York State see:

Talbot et al. Calculating census tract-based life expectancy in New York state: a generalizable approach. Pop Health Metrics, 2018 Jan 26; 16(1):1

Contact: Thomas Talbot, thmstalbot@gmail.com