ACS User Conference, 2017
Alexandria, VA
May 11-12, 2017

Can we map ACS data with "confidence"?

David W Wong
&
Min Sun
George Mason University

Research reported in this presentation was partly supported by the National Institutes of Health (NIH) under Award Number R01HD076020. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Some earlier works were funded under the U.S. Census Bureau contract Order # YA1323-12-SE-0387.
The encouragements and insights from Dr. Nancy Torrieri are greatly appreciated.

WHY MAPS (MAPPING ACS DATA)?

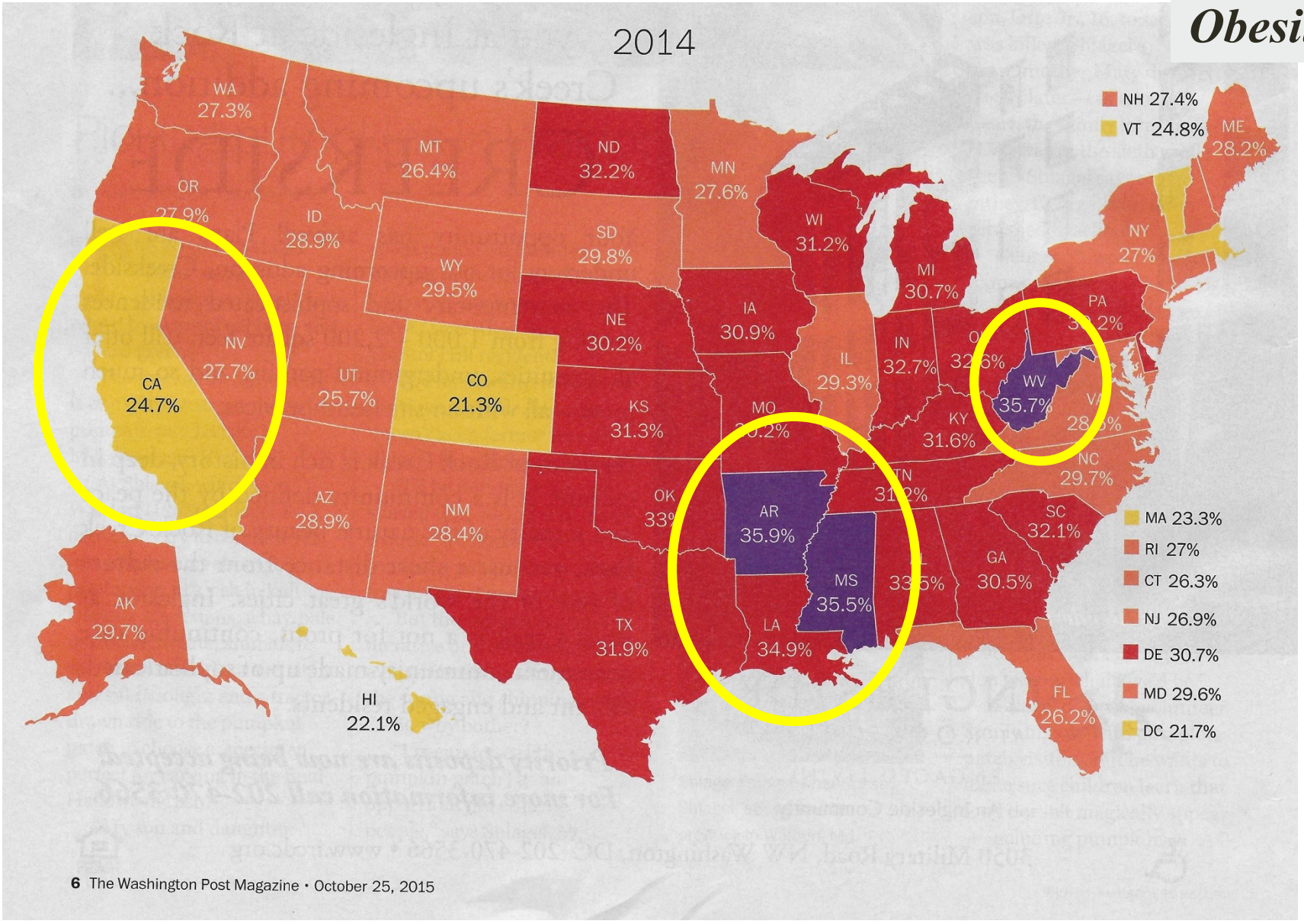
- An "effective"/default means to represent spatial data and results
- Maps are likable in general
- Map-making/GIS – highly accessible
- "A picture speaks a thousand words"
- But Monmonier: "*How to Lie with Maps?*"
- "All maps lie"
- "No map is correct"
- But how much error is acceptable/not too much?

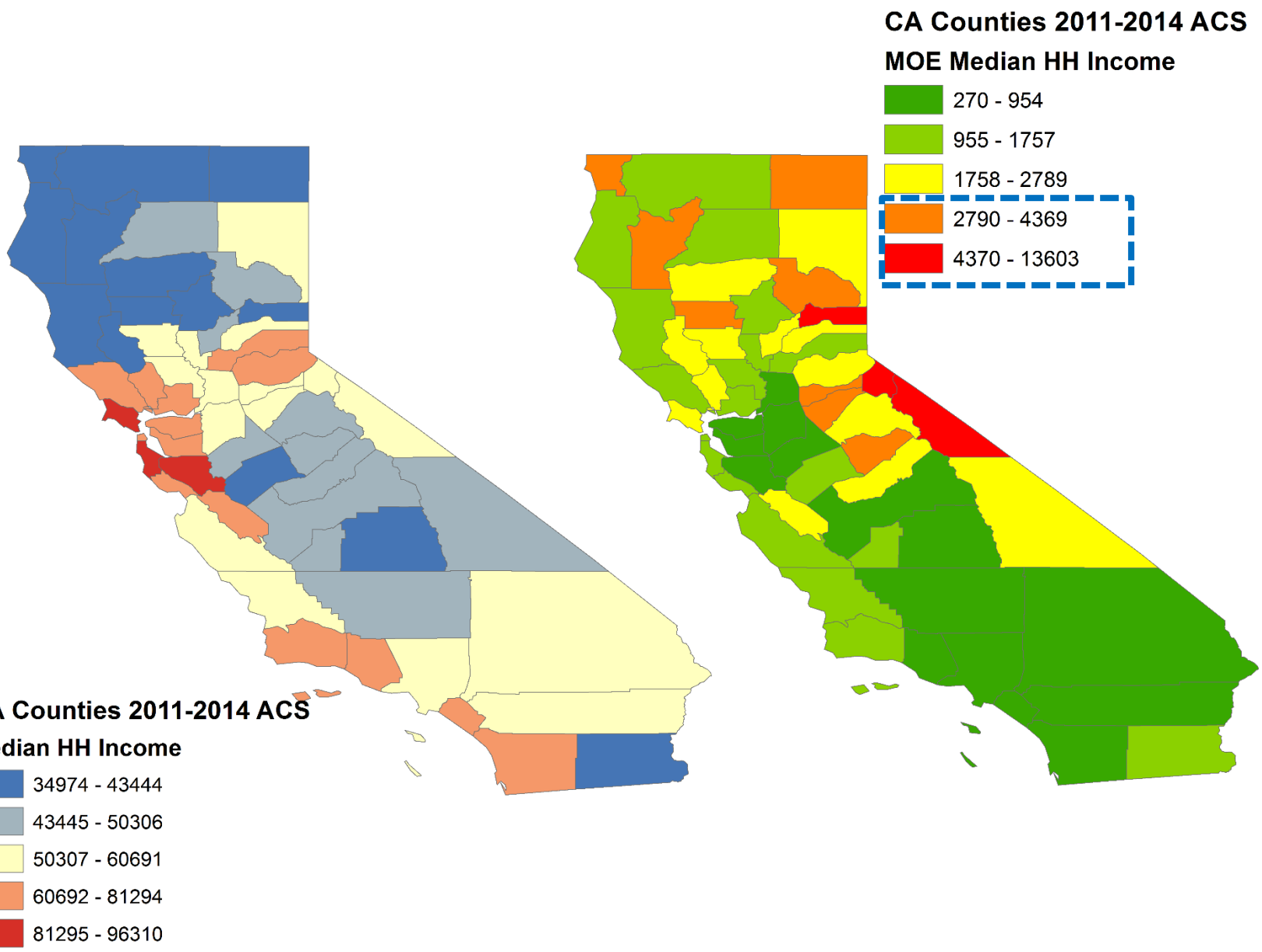
PROBLEMS WITH (CHOROPLETH) MAPPING SURVEY/ACS DATA

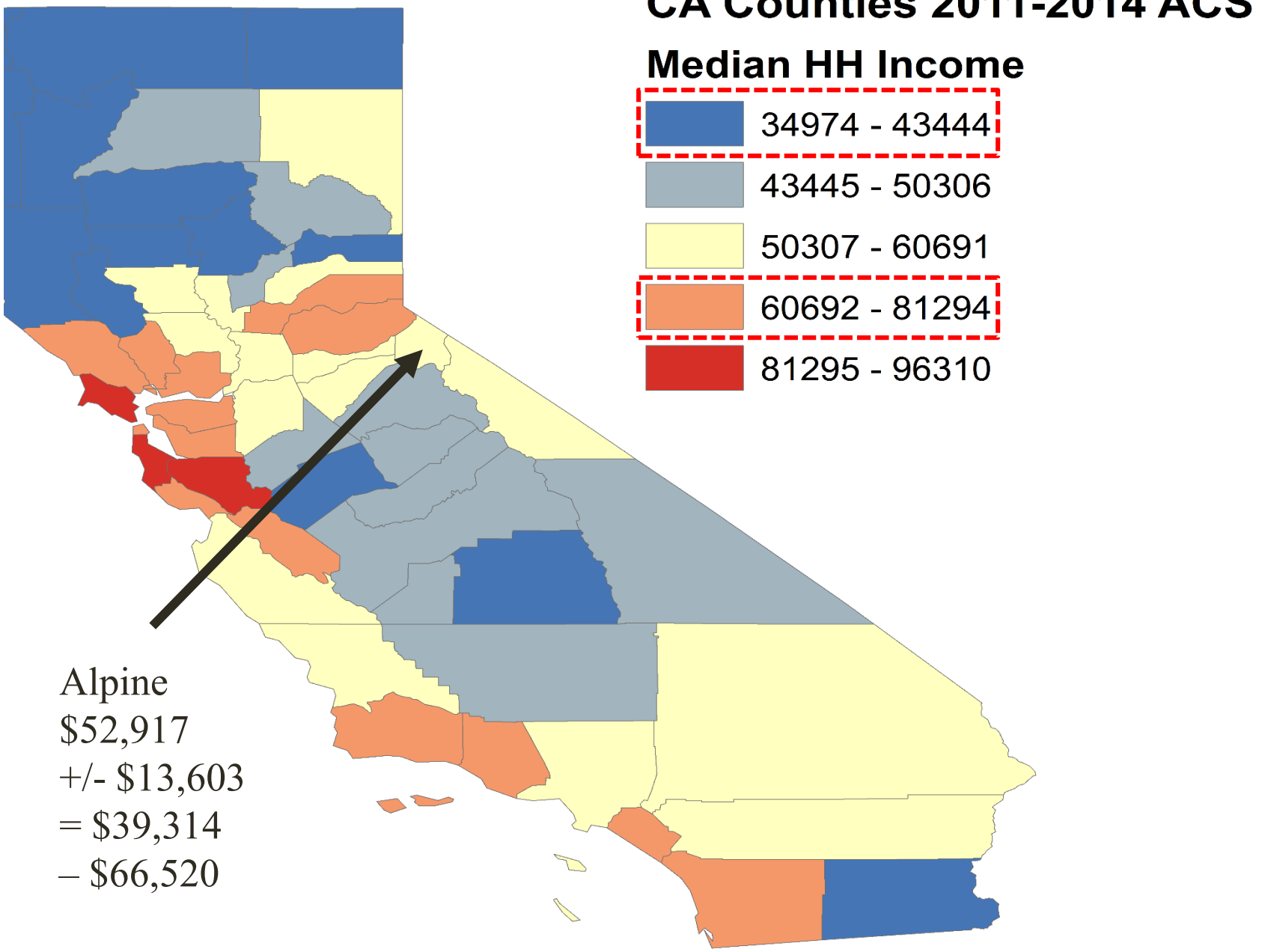
- Sample surveys data: besides ACS, public safety data, SEER, EPA, etc.
- Attribute estimates may have significant levels of uncertainty
- Areal units assigned to different classes are expected to have estimates different from each other
- Spatial patterns are formed because of the systematic differences in estimates.
- But estimates assigned to different classes may not be statistically different (when error is taken into account)
- Spatial patterns may be erroneous (“nothing there”)

ADULT OBESITY: TRUST OF AM HEALTH & R.W. JOHNSON FOUNDATION

State of Obesity

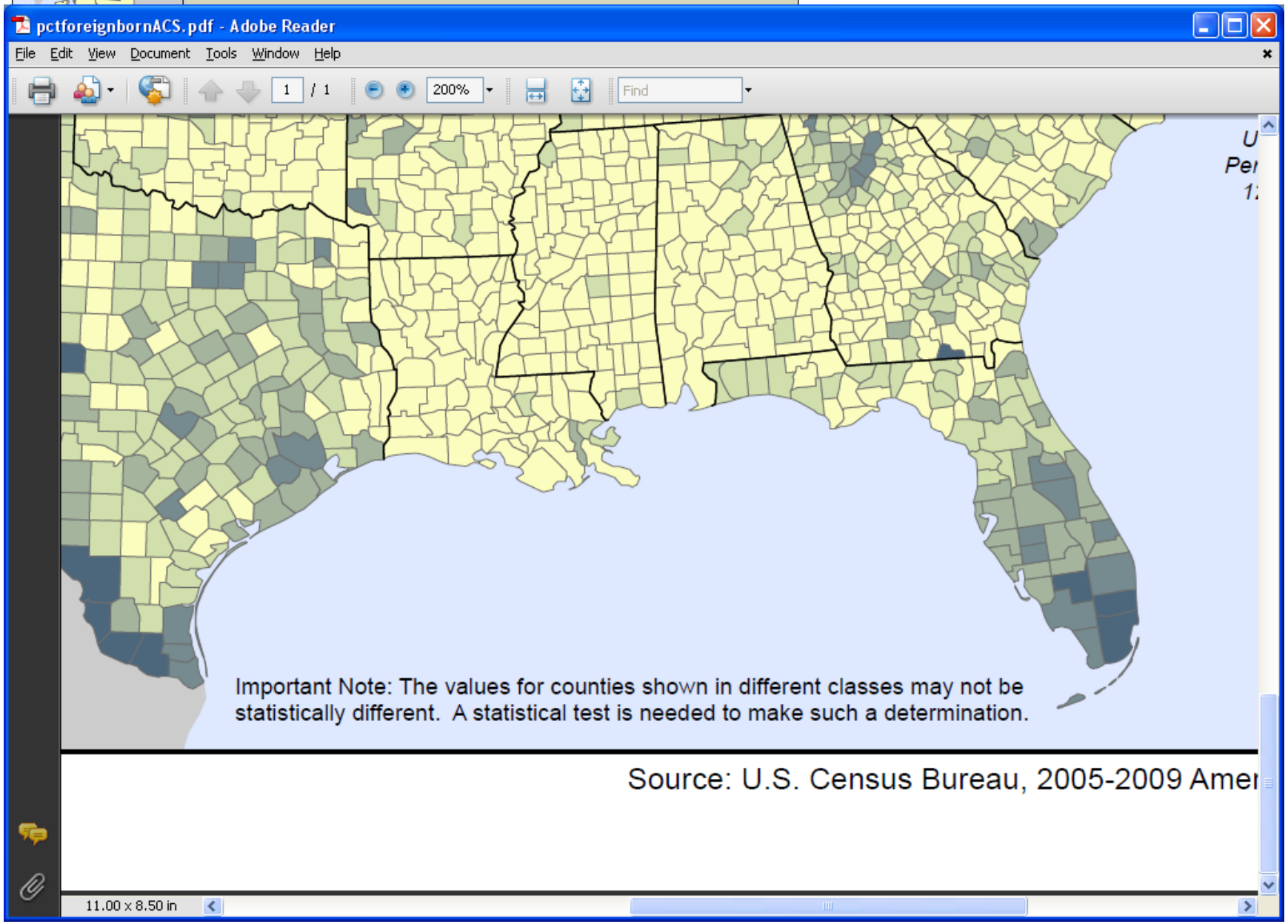



CA Counties 2011-2014 ACS
MOE Median HH Income
270 - 954
955 - 1757
1758 - 2789
2790 - 4369
4370 - 13603
CA Counties 2011-2014 ACS
Median HH Income
34974 - 43444
43445 - 50306
50307 - 60691
60692 - 81294
81295 - 96310

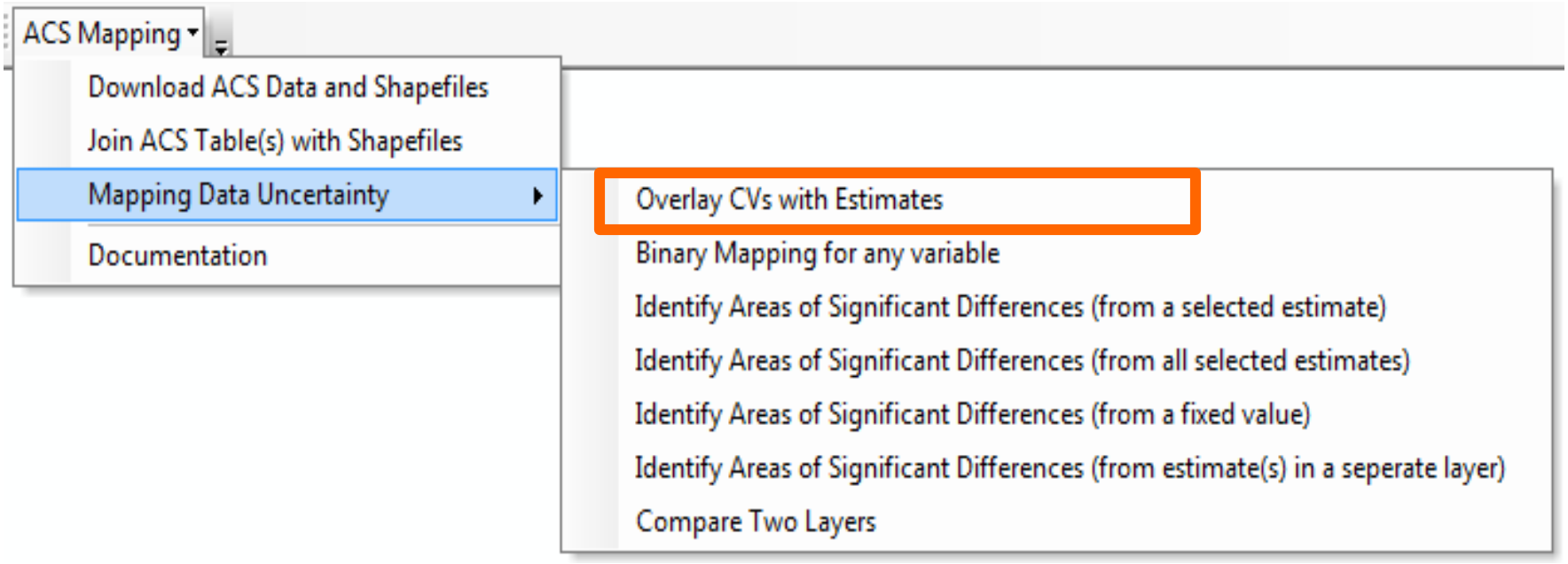

CA Counties 2011-2014 ACS
Median HH Income
34974 - 43444
43445 - 50306
50307 - 60691
60692 - 81294
81295 - 96310
Alpine
$52,917
+/- $13,603
= $39,314
– $66,520

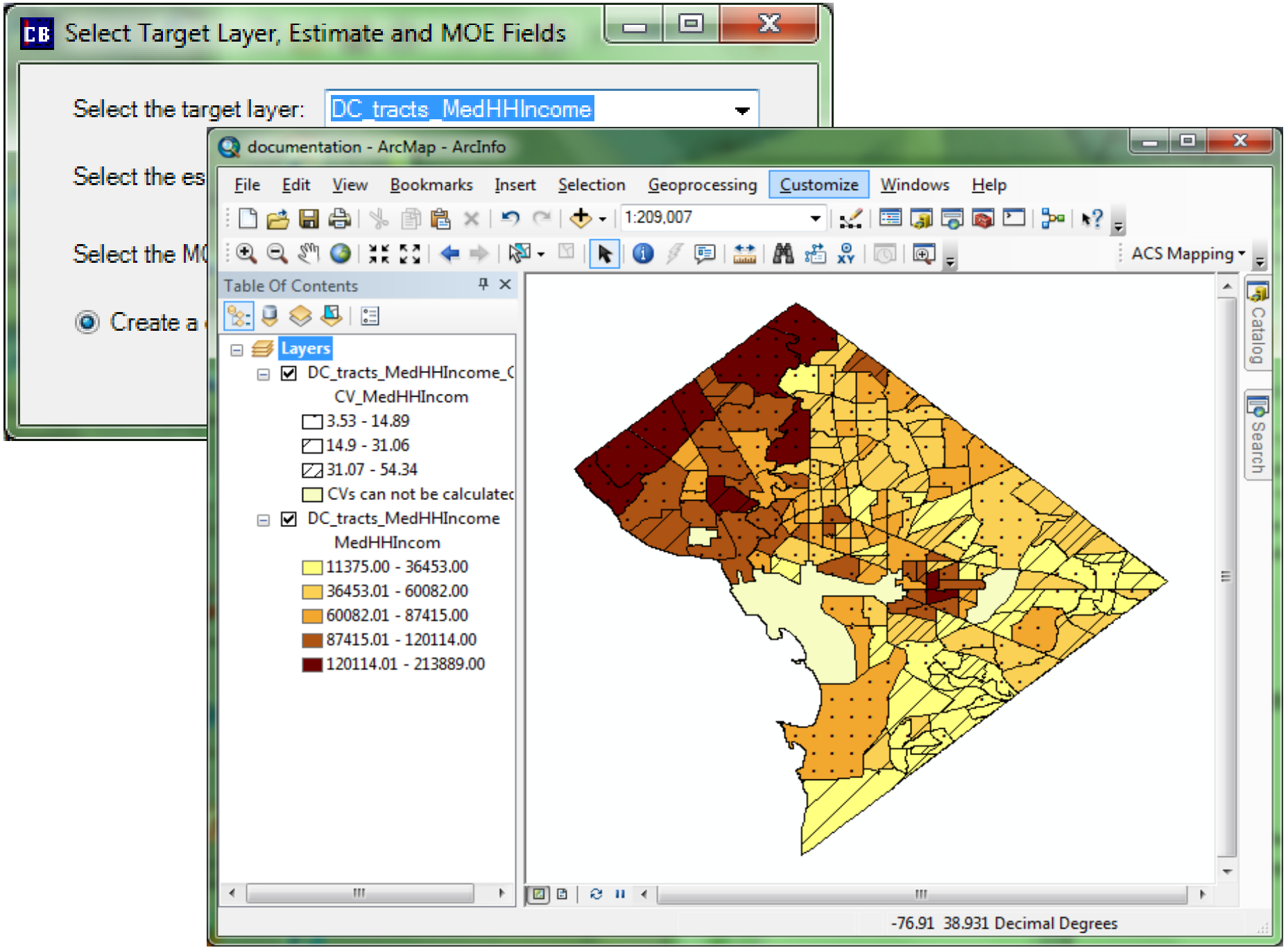
APPROACHES HANDLING/INCORPORATING ERROR IN MAPPING

- Clearly acknowledging unreliability of data


pctforeignbornACS.pdf - Adobe Reader
File Edit View Document Tools Window Help
1 / 1
200%
Find
Important Note: The values for counties shown in different classes may not be statistically different. A statistical test is needed to make such a determination.
Source: U.S. Census Bureau, 2005-2009 Amer
11.00 x 8.50 in

INDIVIDUAL FUNCTIONS OF THE ARCGIS EXTENSION

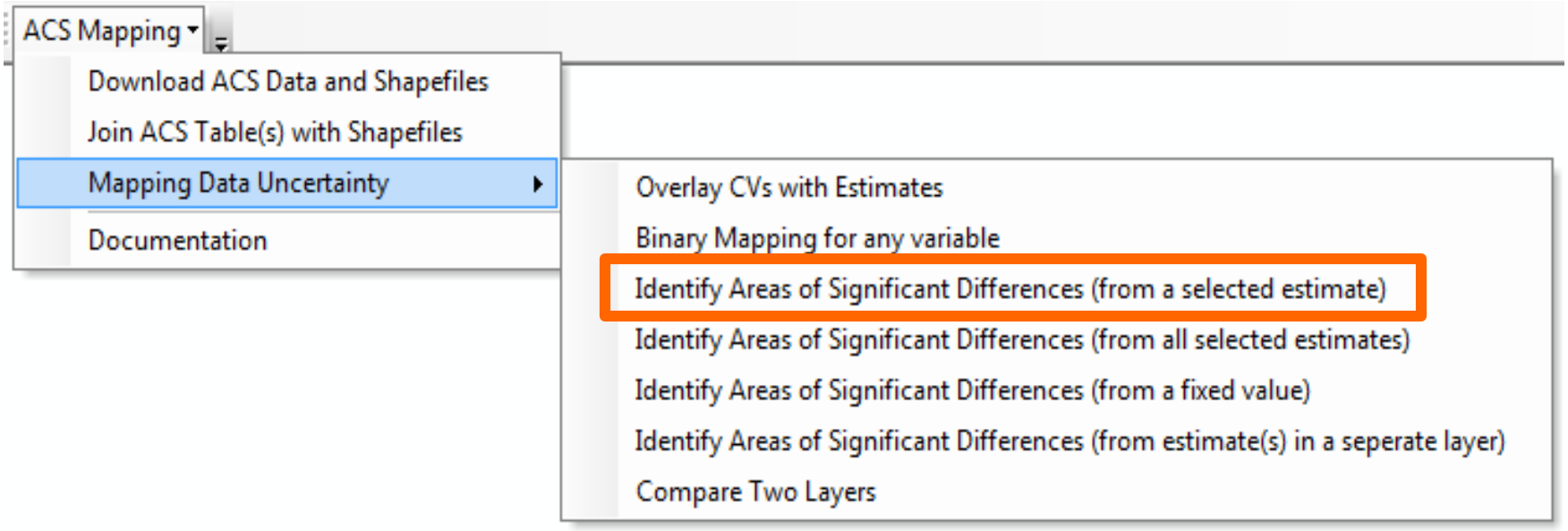



Select Target Layer, Estimate and MOE Fields
Select the target layer:
DC_tracts_MedHHIncome
documentation - ArcMap - ArcInfo
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:209,007
ACS Mapping
Table Of Contents
Layers
DC_tracts_MedHHIncome_C
CV_MedHHIncom
3.53 - 14.89
14.9 - 31.06
31.07 - 54.34
CVs can not be calculatec
DC_tracts_MedHHIncome
MedHHIncom
11375.00 - 36453.00
36453.01 - 60082.00
60082.01 - 87415.00
87415.01 - 120114.00
120114.01 - 213889.00
Catalog
Search
-76.91 38.931 Decimal Degrees

APPROACHES HANDLING/INCORPORATING ERROR IN MAPPING

- Clearly acknowledging unreliability of data
- Help map readers to discern if estimates are different (statistically)

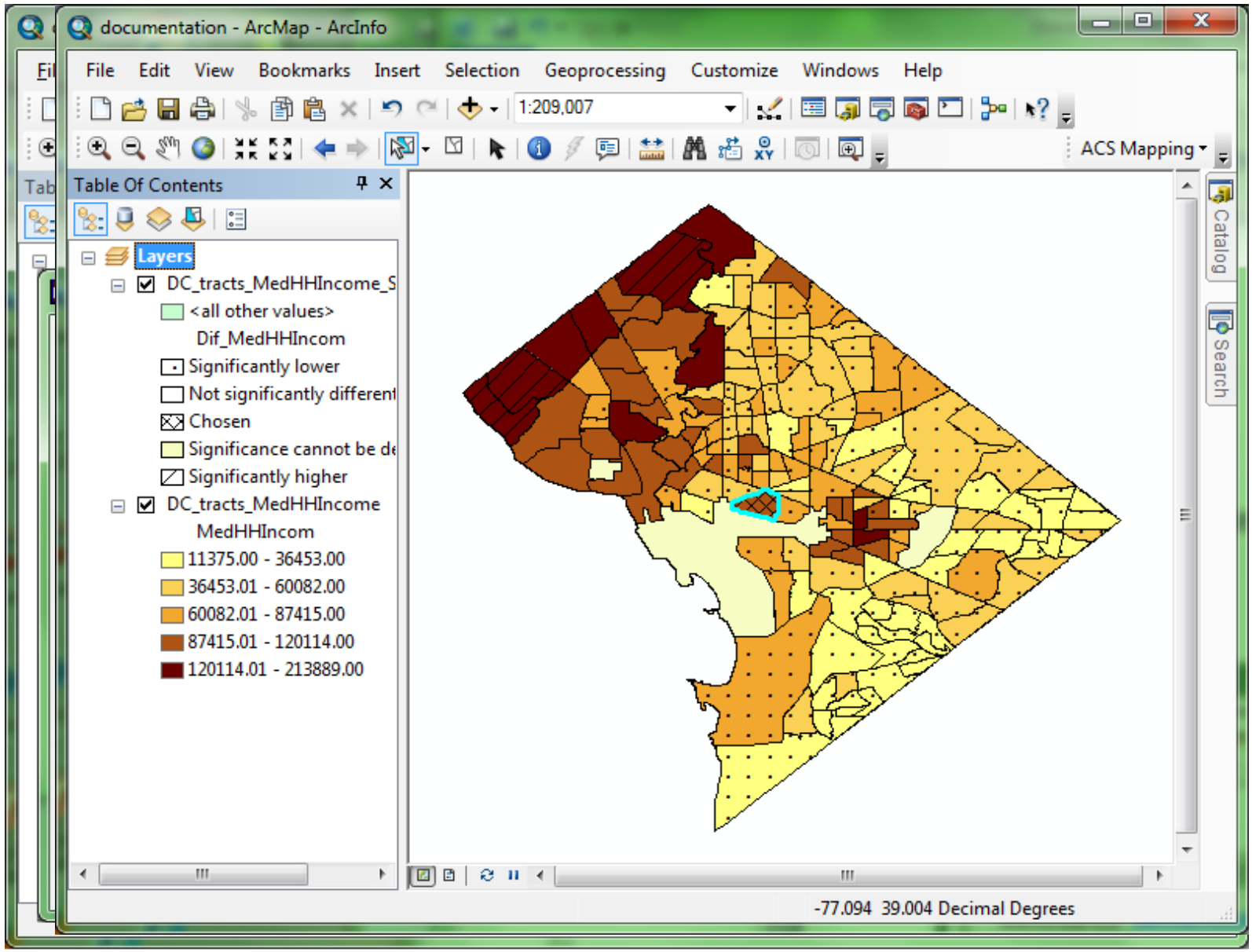
INDIVIDUAL FUNCTIONS OF THE EXTENSION



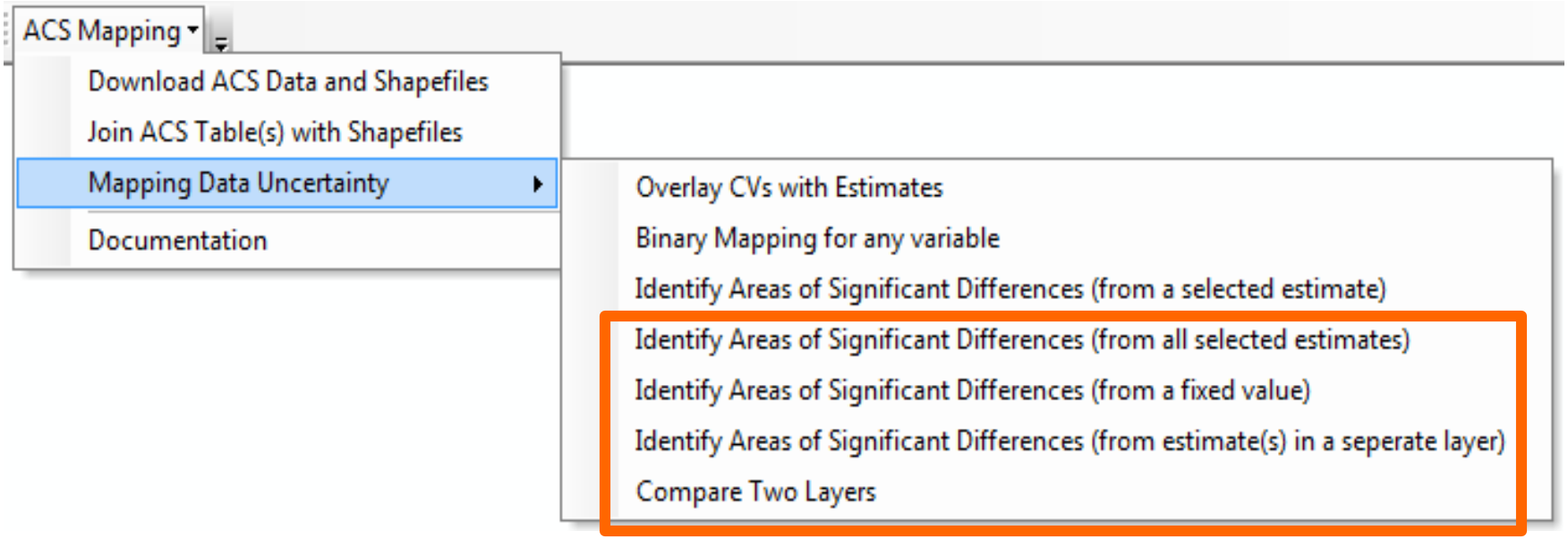
Use the Selection tool to select an areal unit as the reference unit

Compare the estimate of the reference unit with all other estimates

Identify areal units with estimates significant lower and higher than that in the reference unit (at several CLs)



INDIVIDUAL FUNCTIONS OF THE EXTENSION

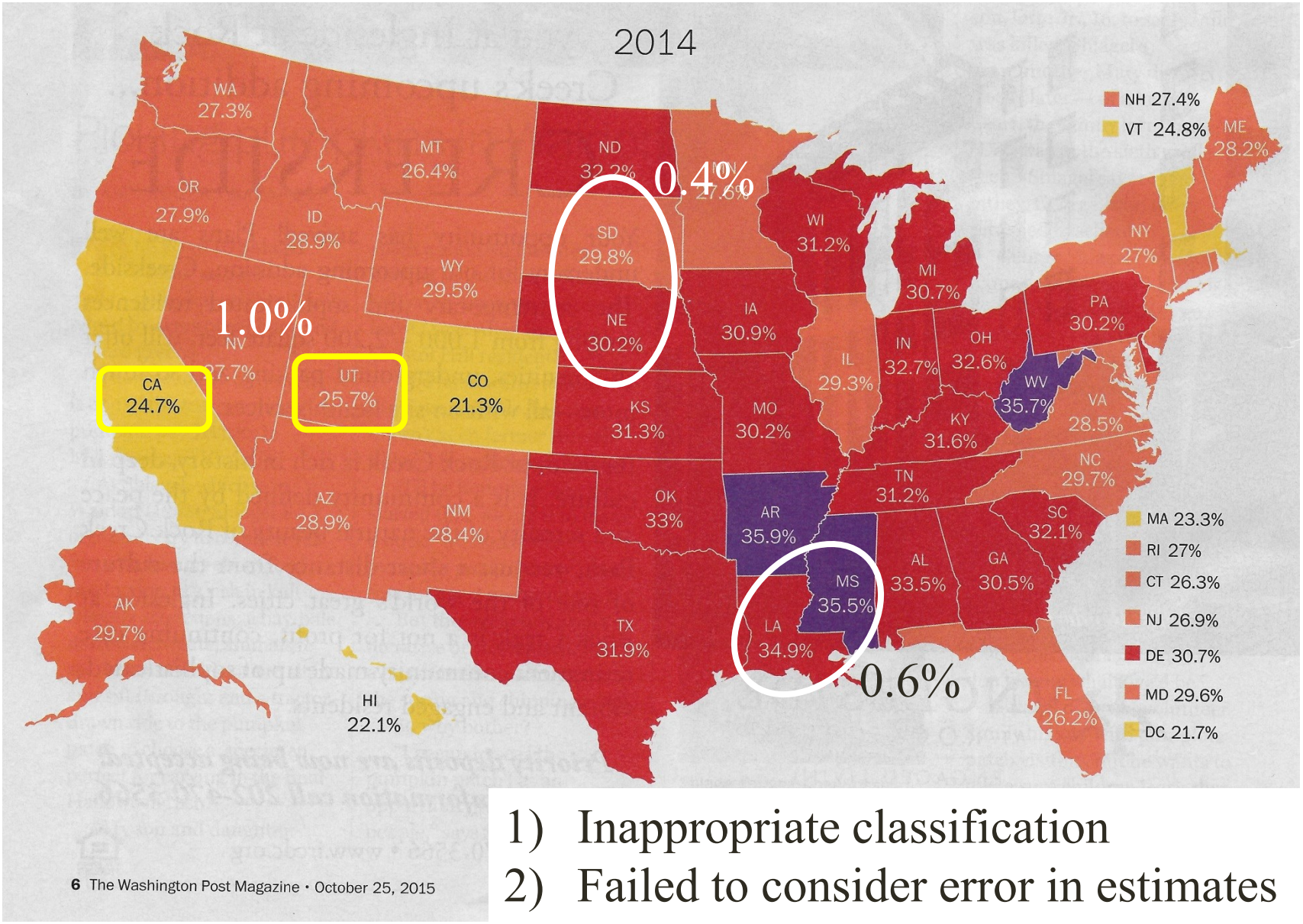


Problems:

- the extension became obsolete – ArcGIS keeps changing every several months

- overlays are not too easy to comprehend

ADULT OBESITY: TRUST OF AM HEALTH & R.W. JOHNSON FOUNDATION



6 The Washington Post Magazine · October 25, 2015

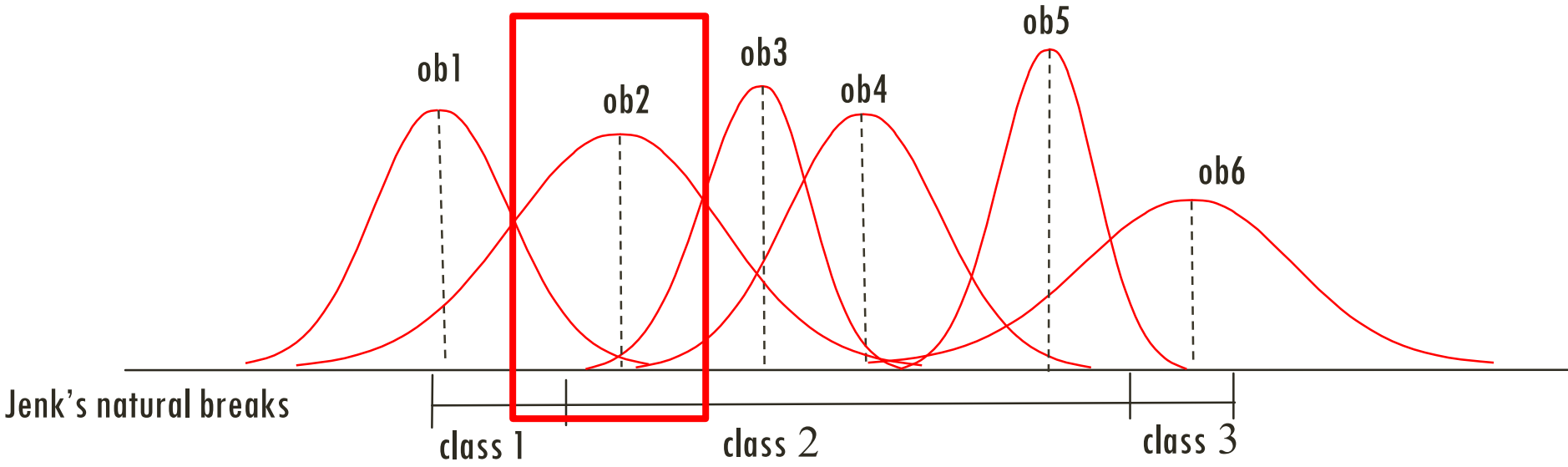
1) Inappropriate classification
2) Failed to consider error in estimates

HANDLING/INCORPORATING ERROR IN MAPPING

- Clearly acknowledging unreliability
- Help map readers to discern if estim
  different (statistically)
- In choropleth maps, when units are assigned to different classes, to what extent are they different?
- In choropleth maps, can we determine class breaks that maximize the differences between classes after considering errors in estimates?

Separability
Classification
Method

PROBLEM CAUSED BY UNRELIABLE ESTIMATES

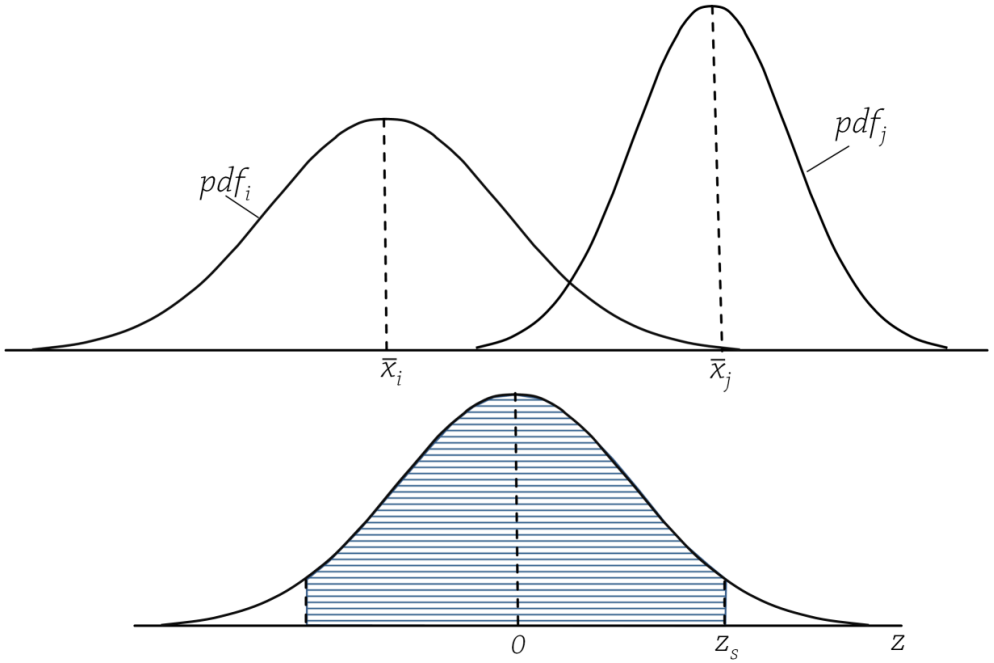


- Observations assigned in one class may have a significant probability of falling into another class (e.g., ob2 to class 1; ob6 to class 2).
- Estimates assigned to different classes may not be really different.
- Spatial patterns presented by the unreliable classification may be misleading!!

MEASURE OF CLASSIFICATION RELIABILITY

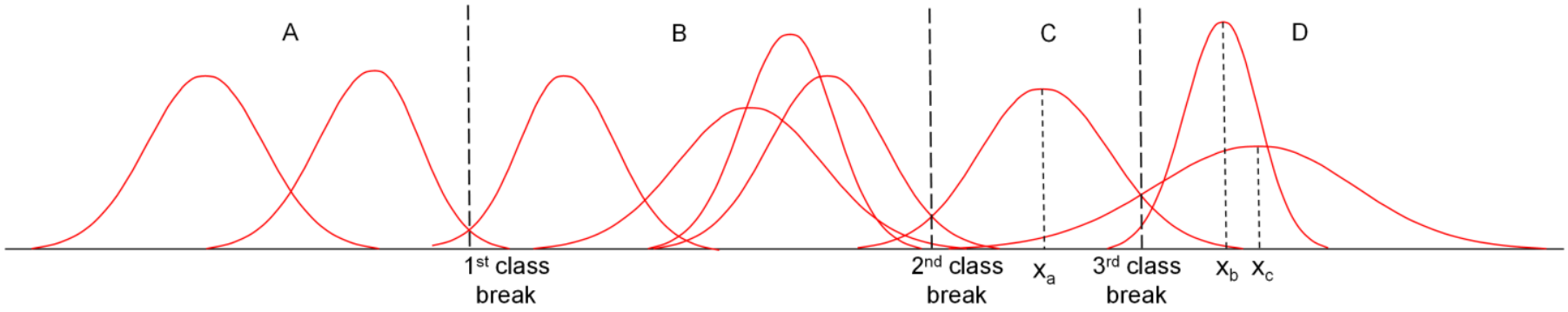
$$CL_{i,j} = \Phi\left(\frac{|\bar{x}_i - \bar{x}_j|}{\sqrt{SE_i^2 + SE_j^2}}\right)$$

Where $|\bar{x}_i - \bar{x}_j|$ is the absolute difference in estimates of two units, SE_i and SE_j are the standard errors of the estimates

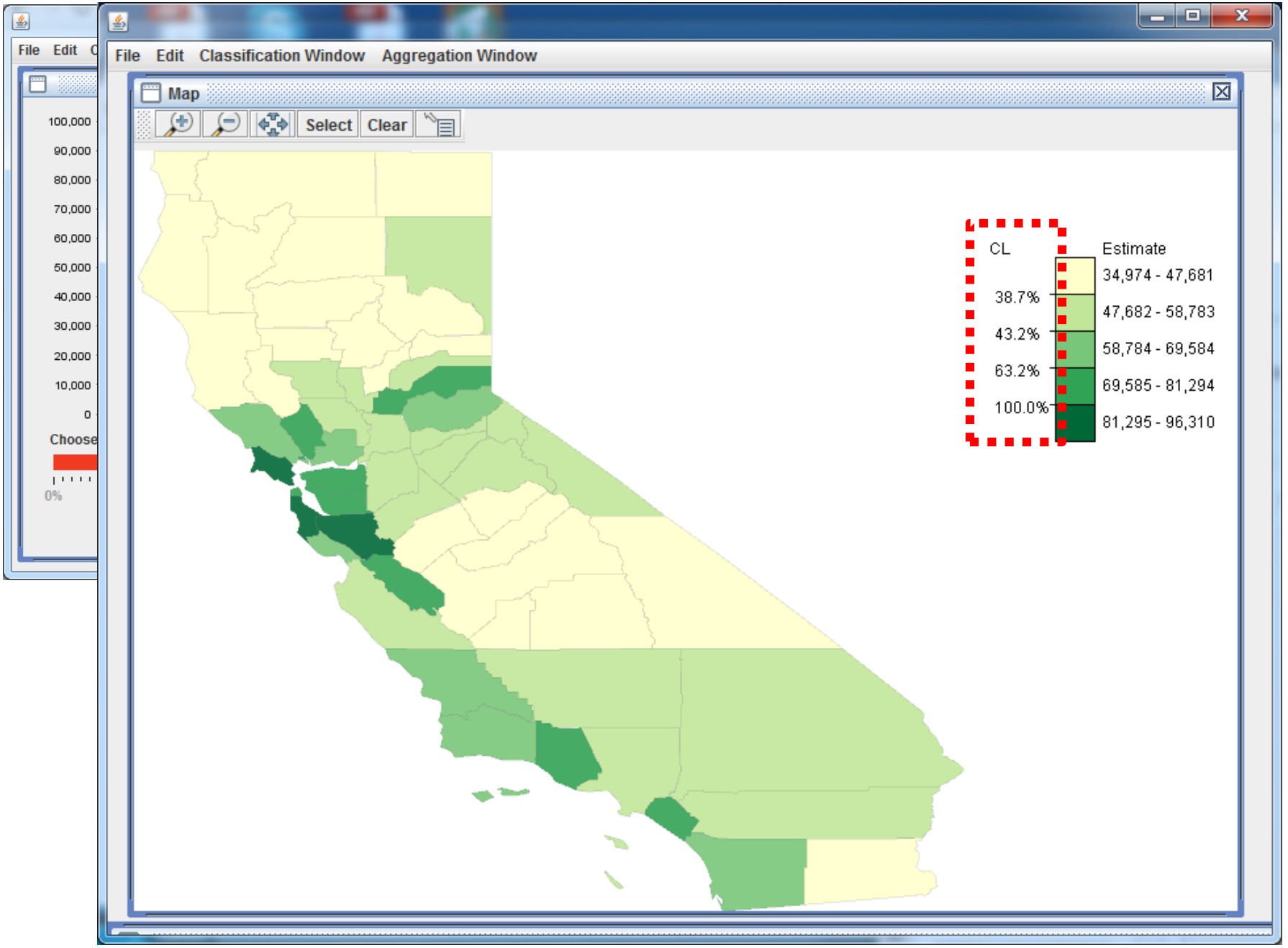


$$S_{A,B} = \min_{i \in A, j \in B} (CL_{i,j}), \qquad i \neq j,$$

$S_{A,B}$ is the minimum probability of difference ($CL_{i,j}$) between a pair of observations i and j in any two different classes A and B.

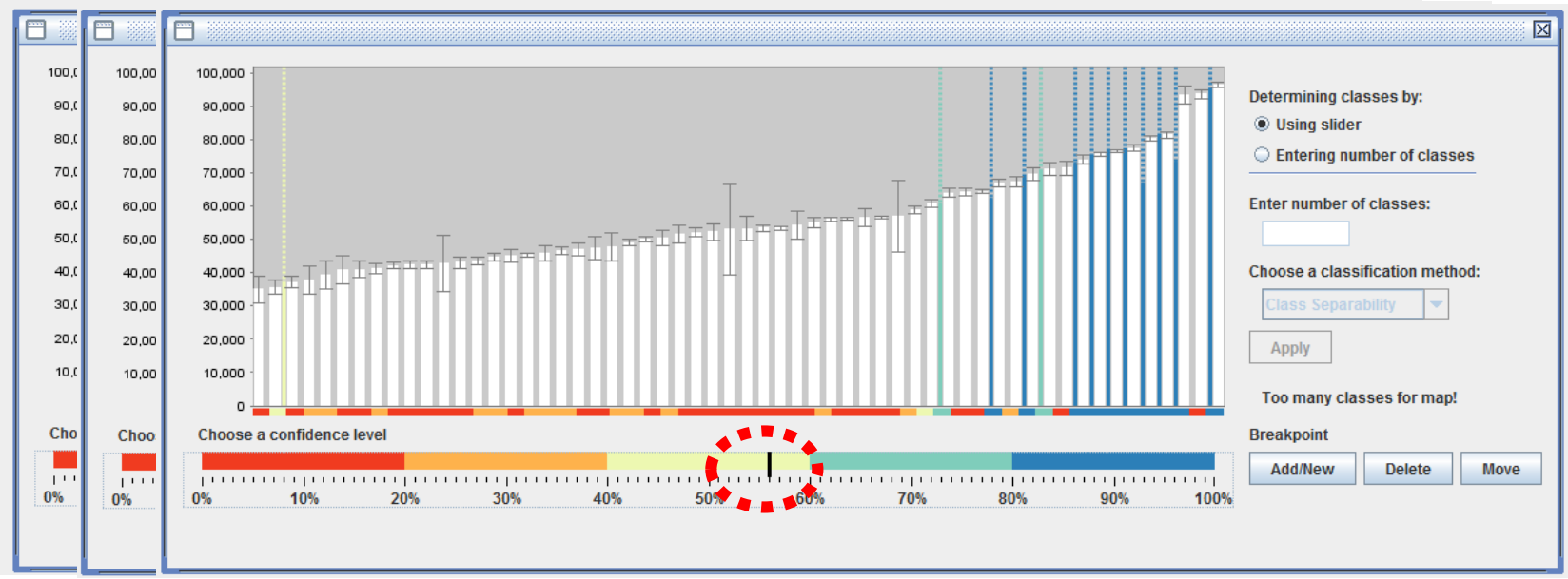


Class Seperability: ~ likelihood that estimates on 2 sides of a break value are different

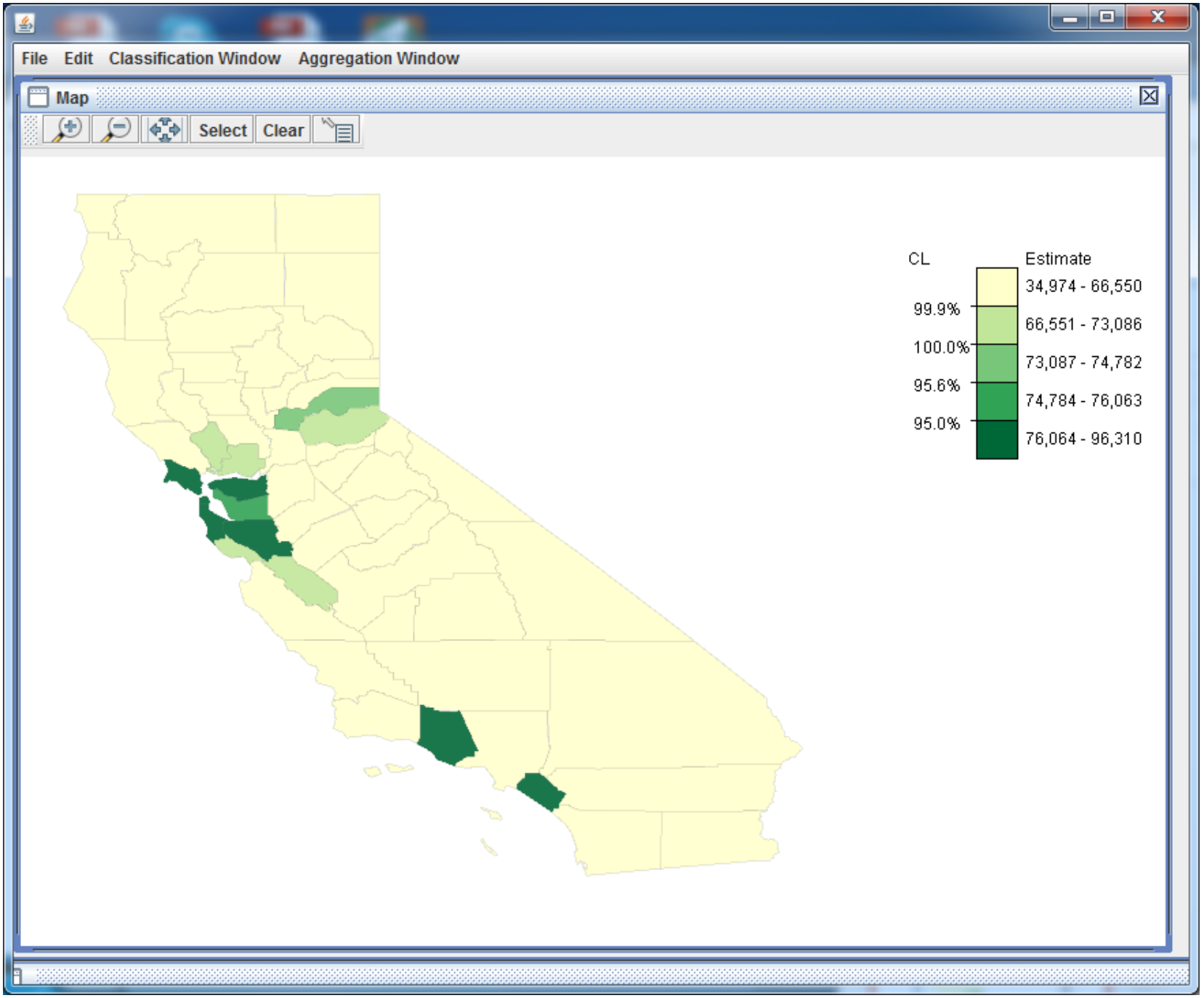

File Edit C
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Choose
0%
File Edit Classification Window Aggregation Window
Map
Select
Clear
CL
Estimate
38.7%
43.2%
63.2%
100.0%
34,974 - 47,681
47,682 - 58,783
58,784 - 69,584
69,585 - 81,294
81,295 - 96,310

DETERMINE CLASSES BY SEPARABILITY

A new classification method - "class separability" : Determine class break values by choosing the break points with the highest $S_{A,B}$ values. (Sun, Wong, and Kronenfeld 2014. A classification method for choropleth maps incorporating data reliability information. *The Professional Geographer*)



High confidence level means fewer classes

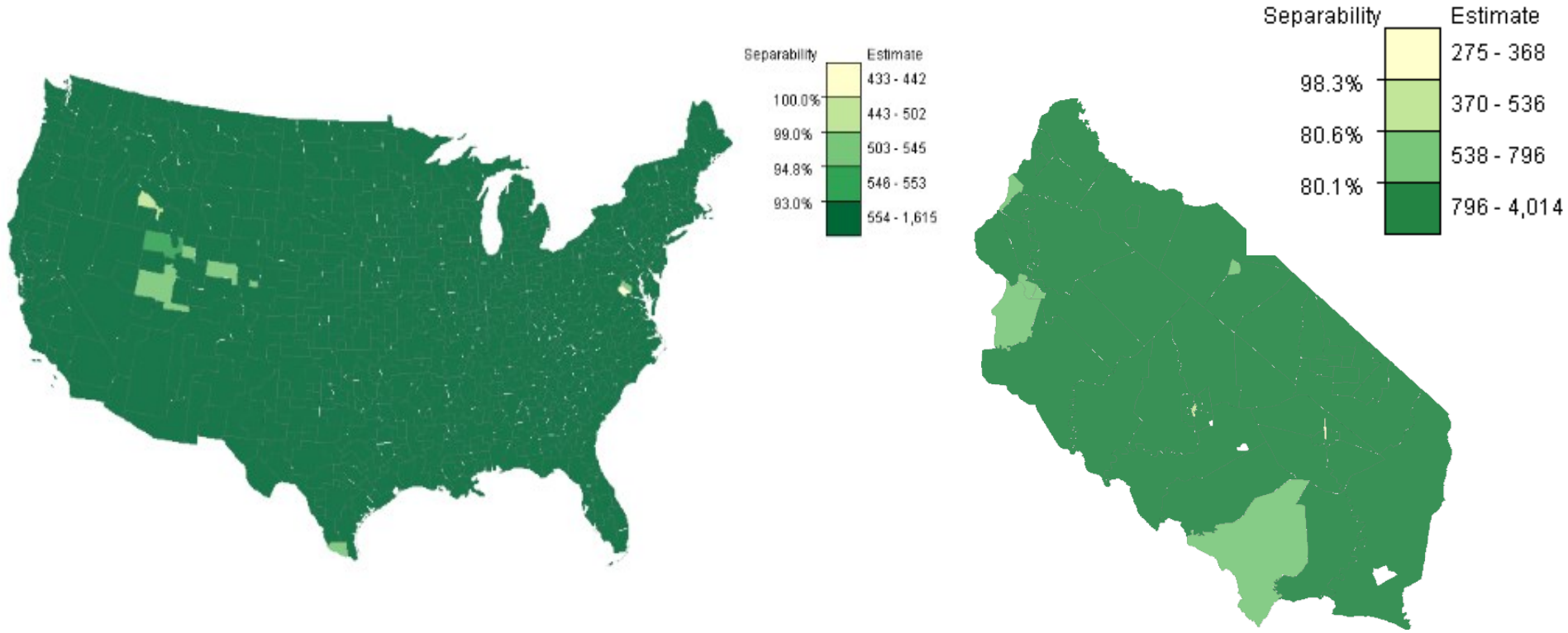

File
Edit
Classification Window
Aggregation Window
Map
Select
Clear
CL
Estimate
99.9%
100.0%
95.6%
95.0%
34,974 - 66,550
66,551 - 73,086
73,087 - 74,782
74,784 - 76,063
76,064 - 96,310

IMPLEMENTATIONS BASED ON THE CLASS SEPARABILITY CRITERION

Hyeongmo Koo, Yongwan Chun, and Daniel A. Griffith. Optimal Map Classification Incorporating Uncertainty Information. *Annals Of The American Association Of Geographers* Vol. 107 , Iss. 3,2017

R. Wei, D. Tong and J. Phillips. An Integrated Classification Scheme for Mapping Estimates and Errors of Estimation from the American Community Survey, *Computers, Environment and Urban Systems*, Volume 63, May 2017, Pages 95–103

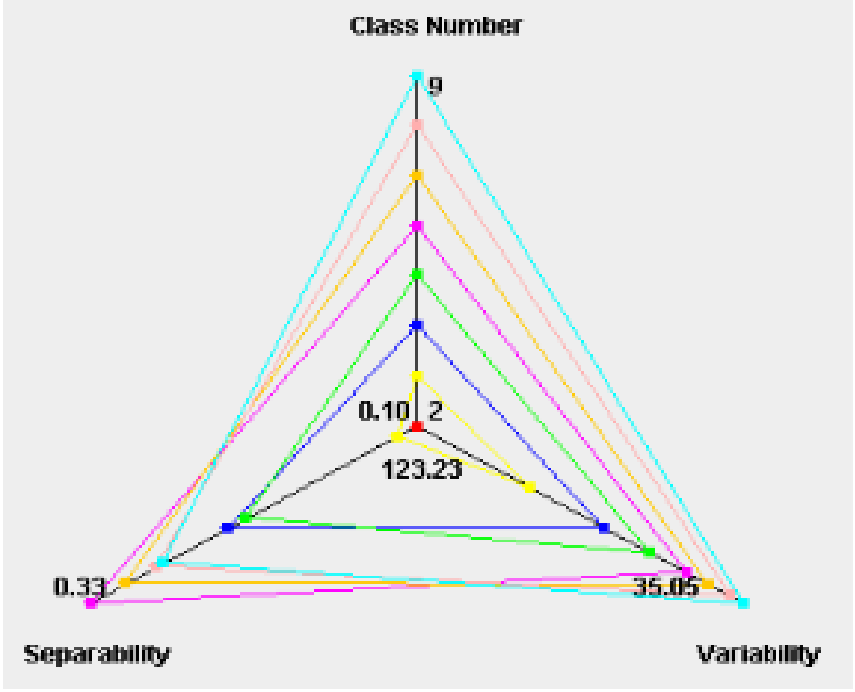
UNBALANCED CLASSIFICATION

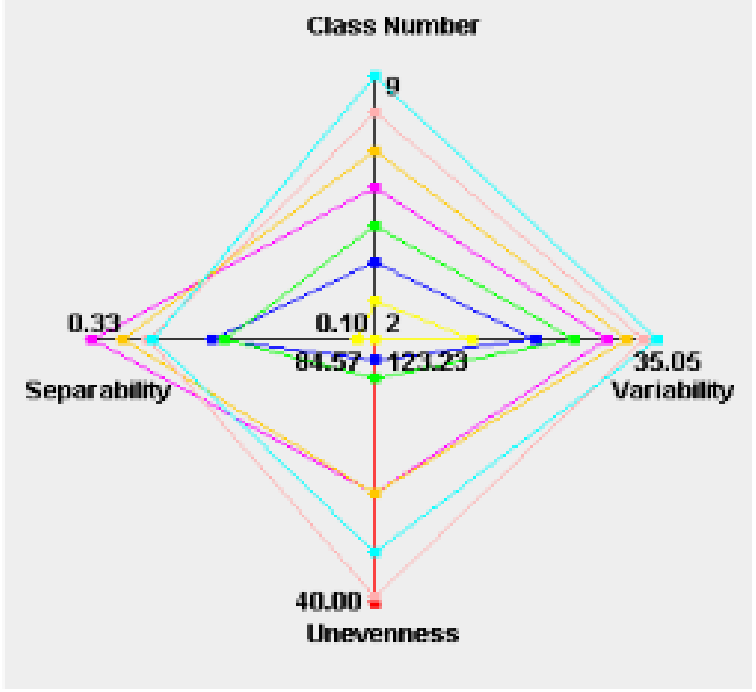


- Need a map with separable, but informative classes
- Solutions:
  - Adjust existing class breaks
  - Determine classification based on criteria in addition to separability

MAPPING BASED ON MULTI-CRITERIA

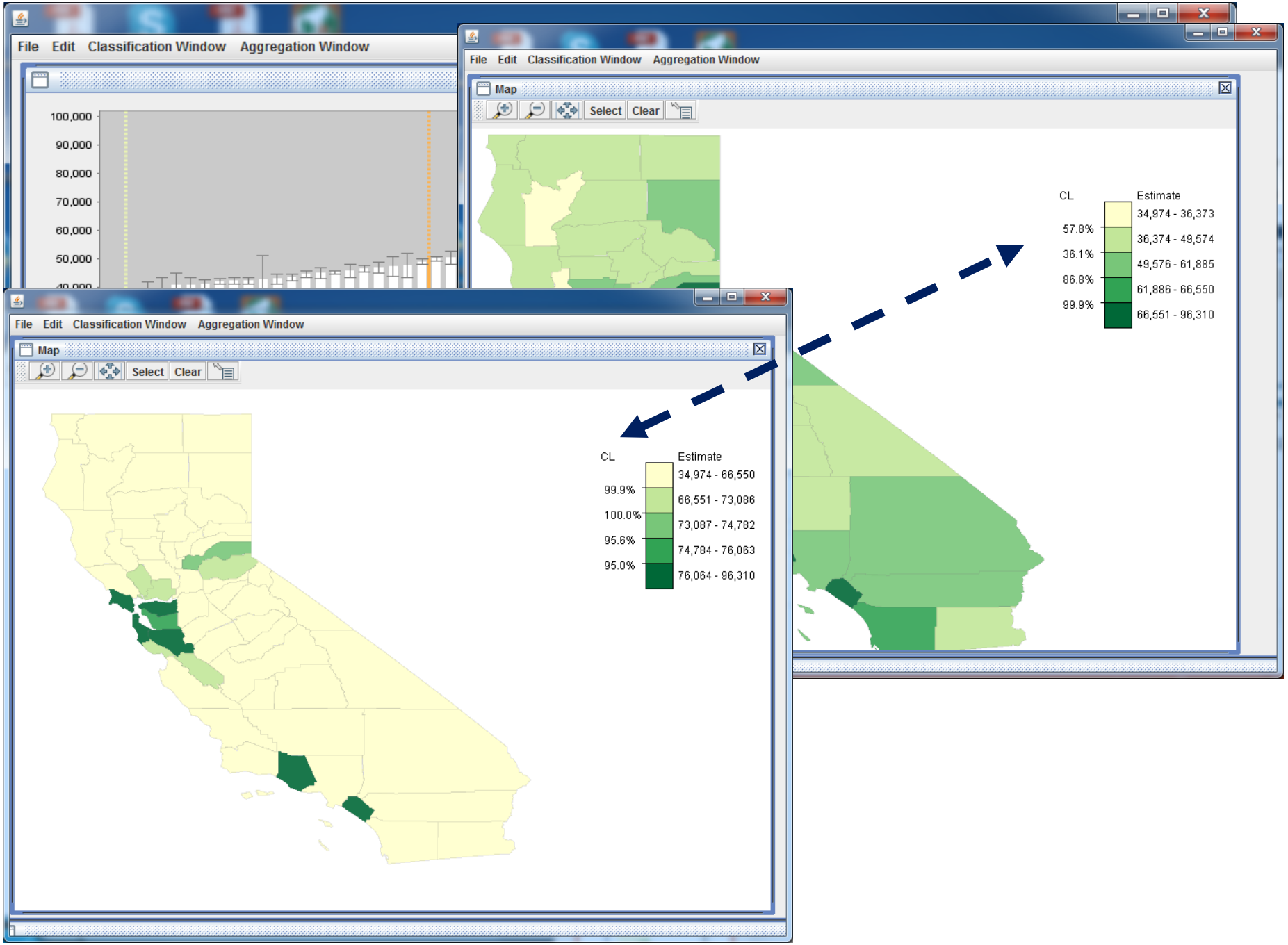
- Involve human intelligence to evaluate the trade-offs among different criteria (including separability)
- Criteria
  - Class separability
  - Number of class (2 to 9)
  - Variability: average within class SD (average)
  - Evenness: distribution of observations across class
- Evaluate the trade-offs and select one scheme



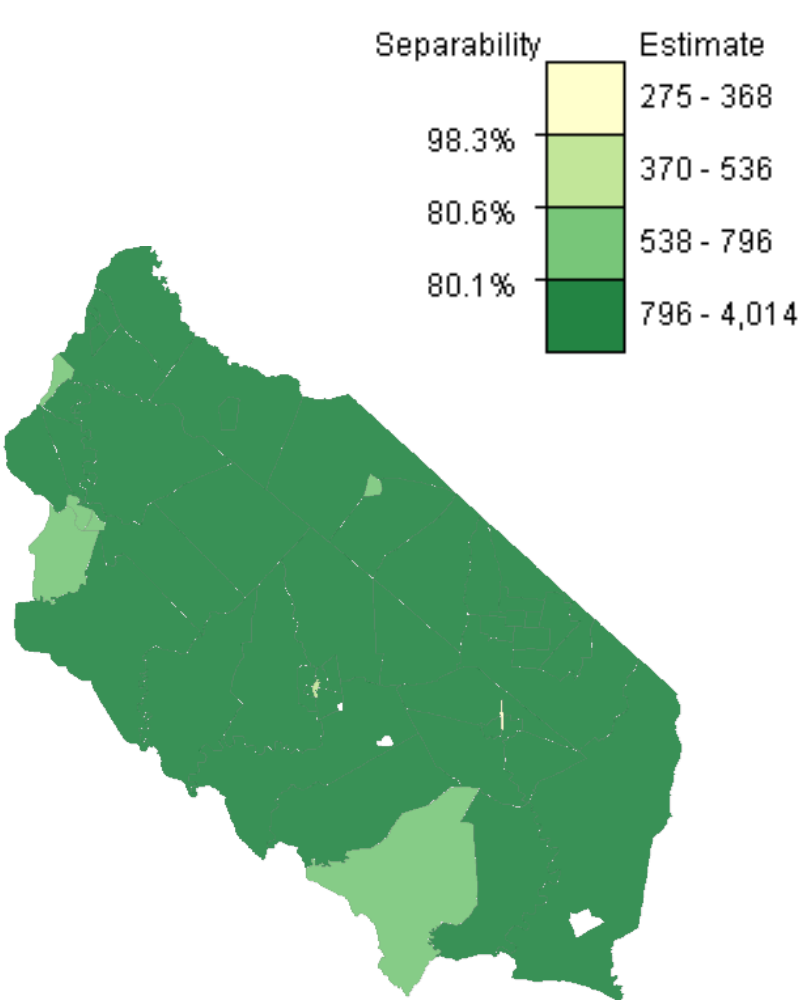


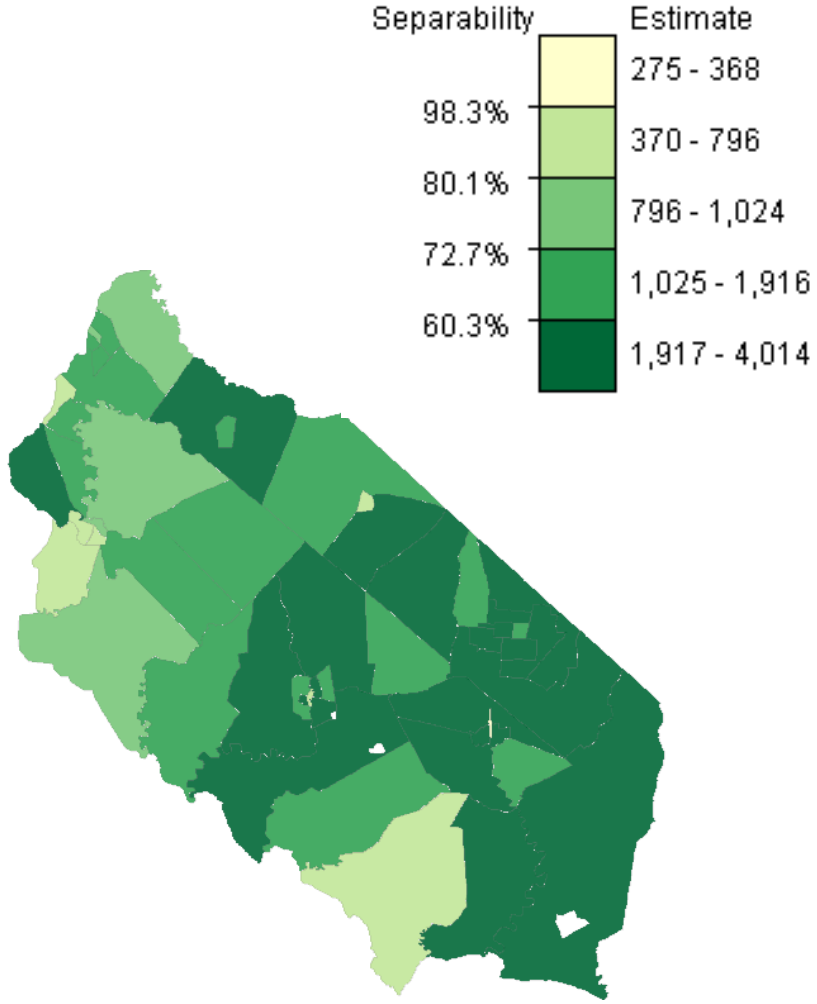
MANIPULATE CLASS BREAKS

- To reduce the unbalance of distribution of estimates across classes, we allow users to manipulate break values
  - If a class has too many estimates, insert a new break value
  - If two break values are too close such that the class in-between has too few observations, remove the class break with lower separability level.


File Edit Classification Window Aggregation Window
100,000
90,000
80,000
70,000
60,000
50,000
File Edit Classification Window Aggregation Window
Map
Select Clear
CL
Estimate
57.8%
34,974 - 36,373
36.1%
36,374 - 49,574
86.8%
49,576 - 61,885
99.9%
61,886 - 66,550
66,551 - 96,310
File Edit Classification Window Aggregation Window
Map
Select Clear
CL
Estimate
99.9%
34,974 - 66,550
100.0%
66,551 - 73,086
95.6%
73,087 - 74,782
95.0%
74,784 - 76,063
76,064 - 96,310

RESULTANT MAP BY ADJUSTING CLASS BREAKS





NCORPORATING ERROR IN

Separability Classification Method

owledging unreliability of data

- In choropleth maps, indicating how a
  different between classes
- In choropleth maps, using classifica
  maximize the differences between
  considering errors in estimates

Interactive Heuristic Spatial Aggregation

- Develop "spatial" methods to make data with relatively large error more usable
- **Bottom Line: making maps that are more informative and accurate (truthful)**

BACKGROUND (1)

Needs a way to reduce error and make data more usable

- Increasing sample size can reduce the SE, raising the estimate reliability, and making the estimates more usable.
- Creating new larger units with larger sample sizes by merging units - **Spatial Aggregation**

BACKGROUND (2)

Automated optimization algorithms: undesirable consequences (costs)

- Units with reasonable quality estimates are subject to the "risk" of being aggregated, changing the geography of units that may not need to be changed.
- may be difficult to incorporate the user's local knowledge of the study area, recognizing the presence of neighbourhoods and taking the local boundaries into account during the aggregation process.

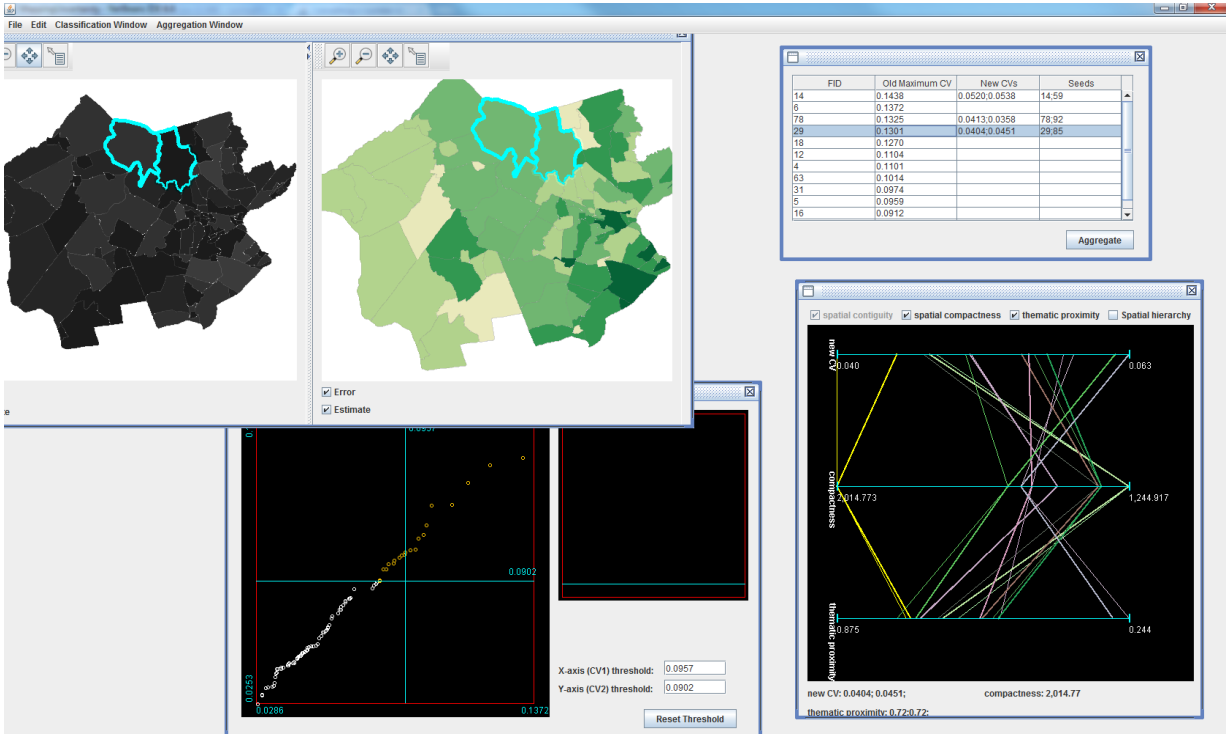
OBJECTIVES

Develop a "new" zoning system, which

- Suppresses error to a level acceptable to the user.
- Resembles the original zonal system as much as possible.
- May incorporate the user's local knowledge of the study area
- Can incorporate more than one variables

GEOVISUAL ANALYTICAL TOOLKITS

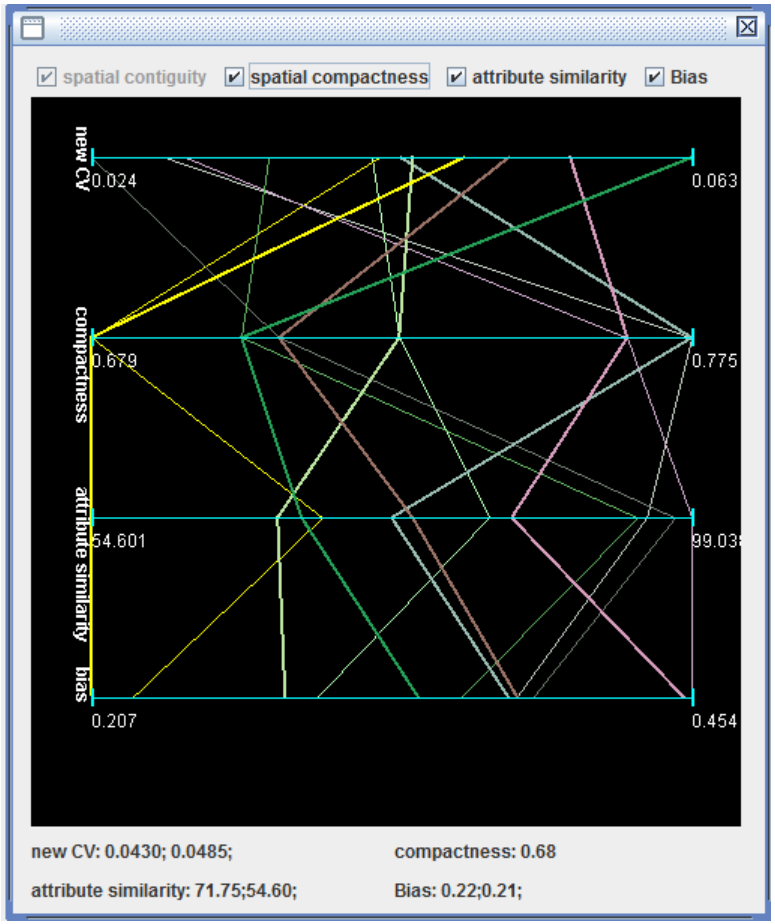
- Statistical plots
- Maps
- Interactive graphic elements to capture user's inputs
- Linked graphics (console/table, plots, and maps)



PARALLEL PLOT

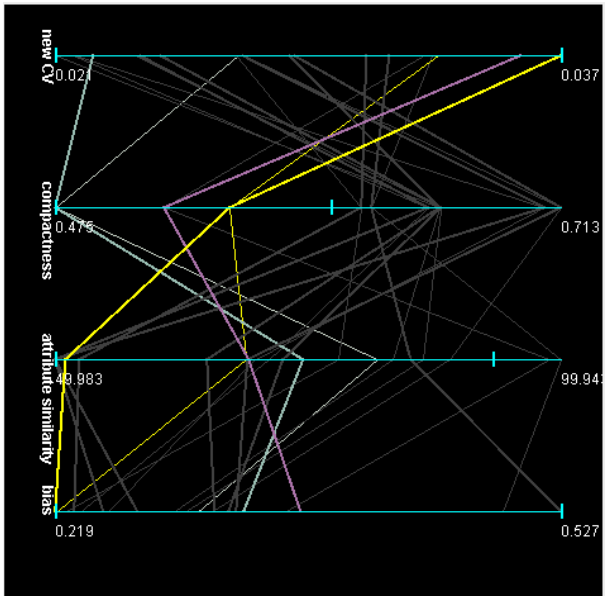
For each seed (unit that needs to be aggregated)

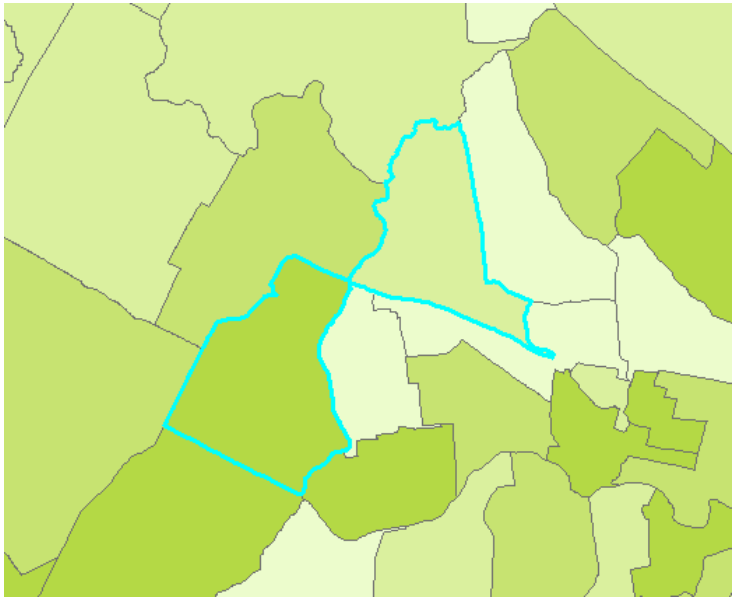
- Axis: criterion - desirable values are aligned to the left
- A set of line segment: a candidate evaluated by different criteria
- Color: Separate different candidates
- Click to select one candidate as the most desirable



MAPS

- Display the locations of seeds and aggregation candidates
- Primarily used to evaluate the compactness/shape of areal units
- Also allow users to consider local/neighborhood knowledge



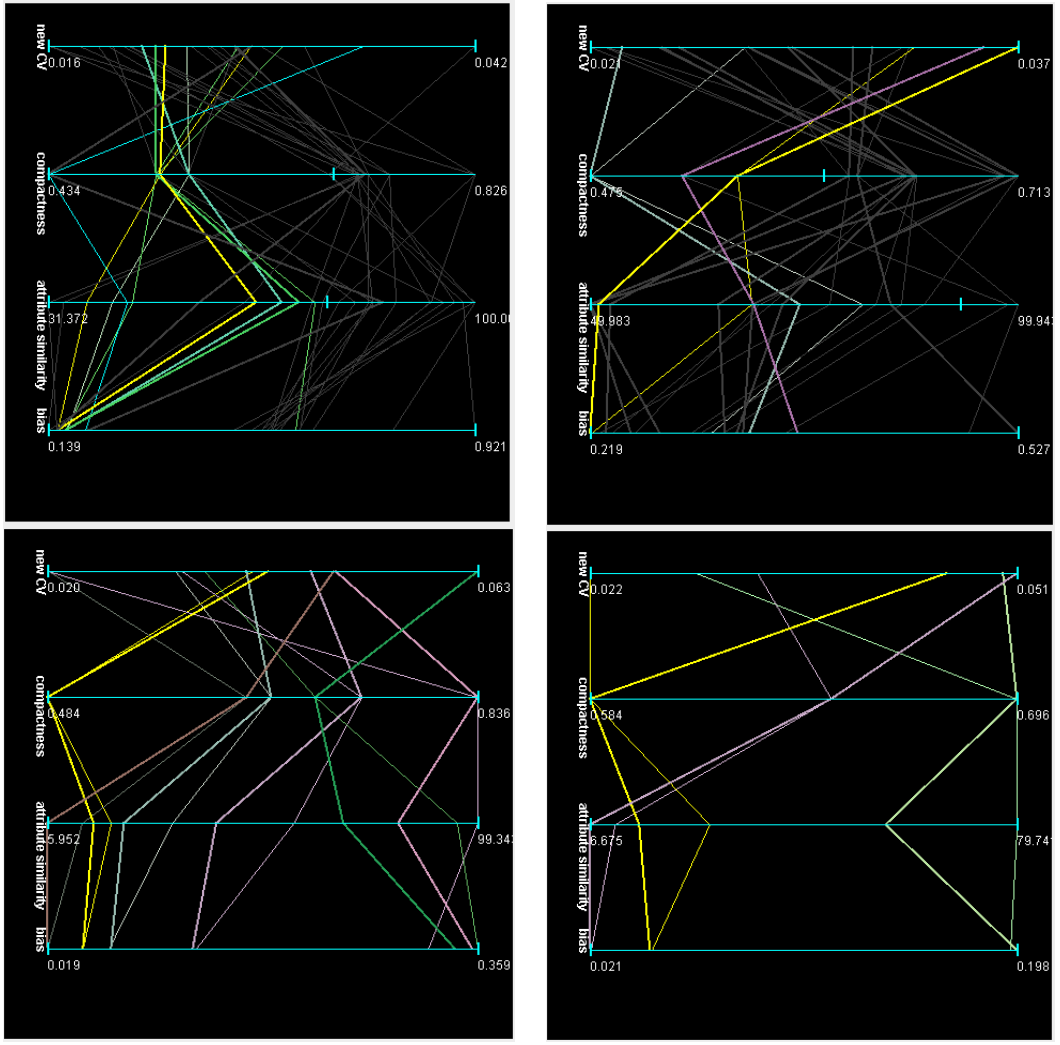


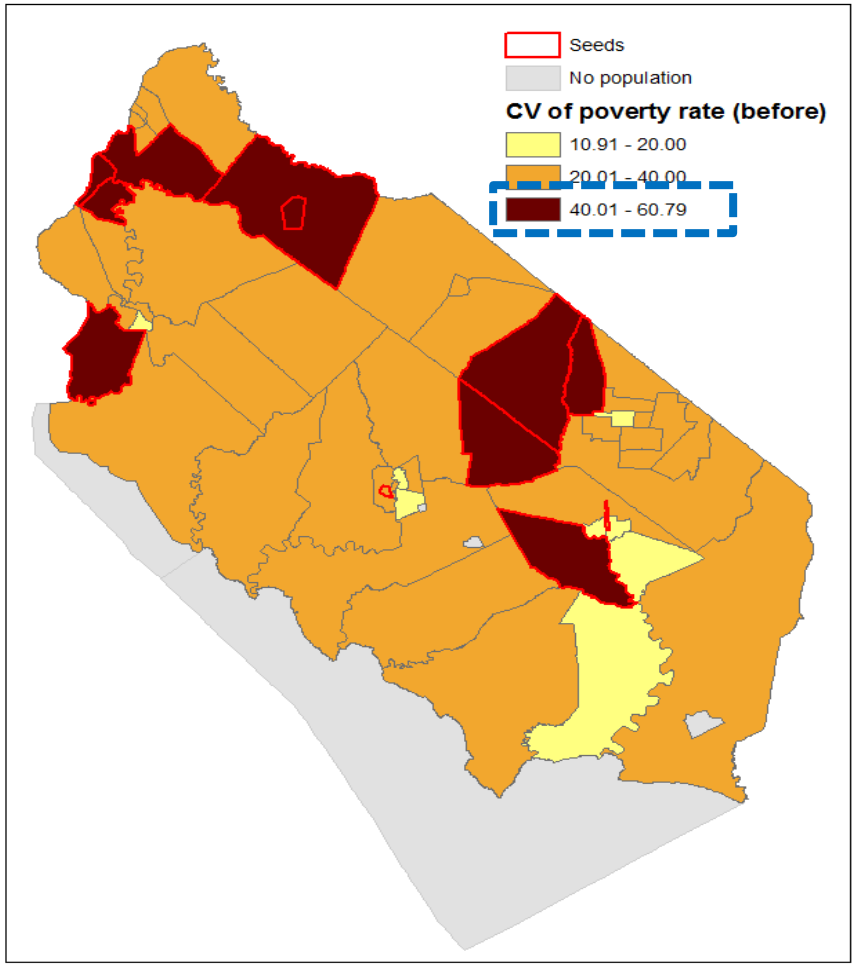
DESIRABLE RESULTS

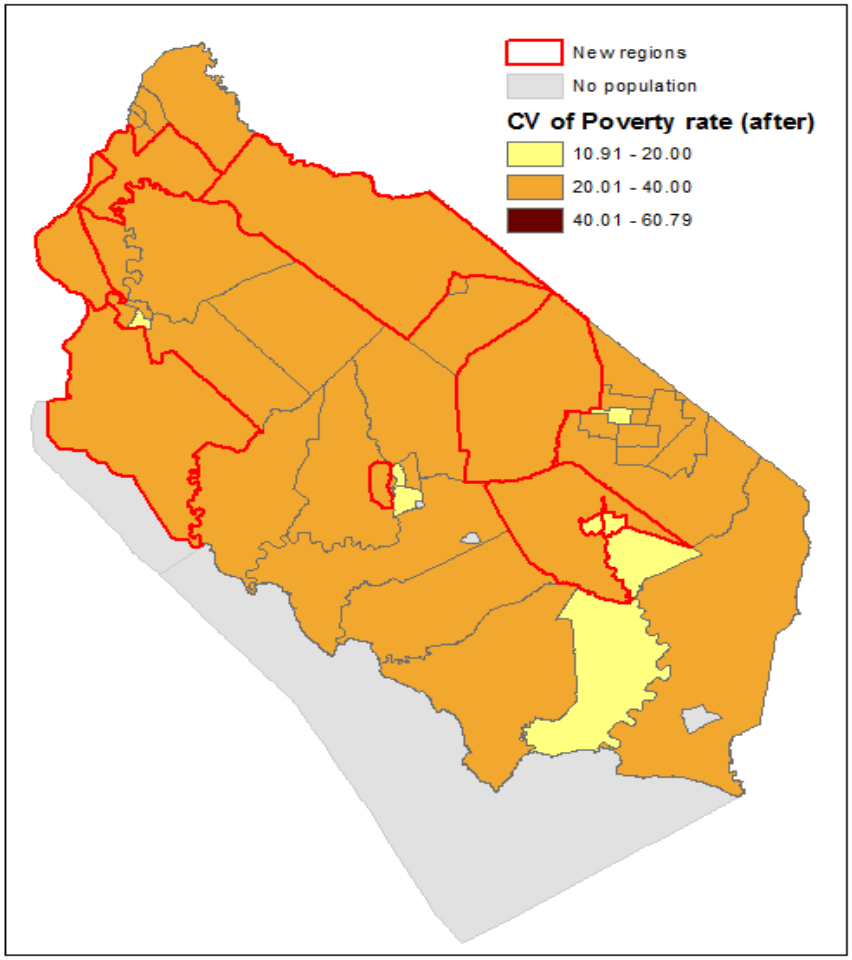
We recommend: The most desirable scheme should be the one with “good or moderate” values in all criteria.

Lower weight on error (all candidates meet the threshold criterion)

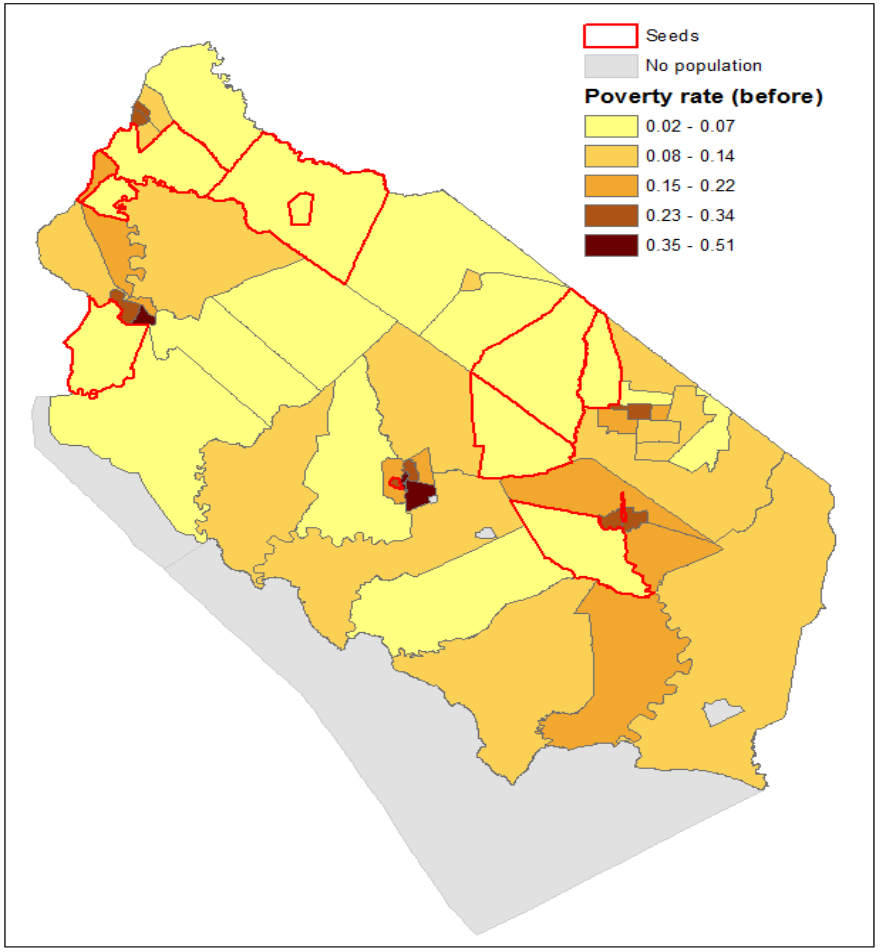
Higher weight on bias

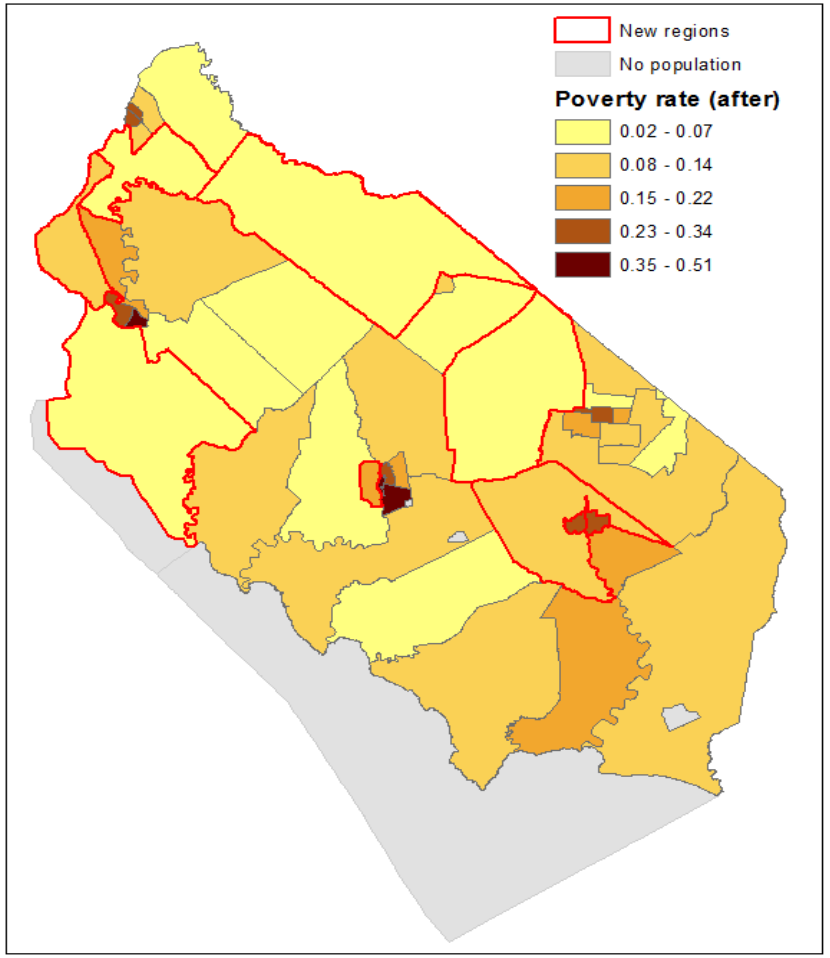






Maps for the CV of poverty rate estimates before (left) and after (right) aggregation with seeds and new zone highlighted (classification method: manual)





Maps for the poverty rate estimates before (left) and after (right) aggregation with seeds and new zone highlighted (The map on the left is made by Jenk's natural breaks method and the map on the right uses the same class break values to facilitate comparison.)

SUMMARY (1): HOW TO IMPROVE MAPPING OF ACS DATA?

- Acknowledge the reliability of ACS estimates
- In choropleth mapping:
  - letting users to interactively compare if values in different units are statistically different
  - determine the likelihood that values in different classes are statistically different
  - creating class breaks that maximize the differences between classes after considering errors in estimates

SUMMARY (2): HOW TO IMPROVE MAPPING OF ACS DATA?

- In choropleth mapping:
  - Allowing users to determine the separability levels between classes but considering other classification criteria
- Develop an interactive spatial aggregation framework to reduce the error levels of estimate to make ACS data more usable
- ***Bottom Line: making maps that are more informative and accurate (truthful)***

THANKS YOU! Q&A

Software download: http://geospatial.gmu.edu/

References:

Sun, M. and Wong, D. W. S. 2010. Incorporating data quality information in mapping the American Community Survey data. ***Cartography and Geographic Information Science*** 37 (4): 285-300.

Wong, D. W. and M. Sun. 2013. Handling data quality information of survey data in GIS: A case of using the American Community Survey data. ***Spatial Demography*** 1(1): 3-16.

Sun, M., Wong, D. W. and Kronenfeld, B. J. 2014. A classification method for choropleth maps incorporating data reliability information. ***The Professional Geographer***. 67(1): 72-83. DOI=10.1080/00330124.2014.888627

Sun, M., D. W. Wong, and B. Kronenfeld. 2016. A heuristic multi-criteria classification approach incorporating data quality information for choropleth mapping. ***Cartography and Geographic Information Science*** DOI: 10.1080/15230406.2016.1145072

Sun, M. and D. W. Wong. Spatial aggregation as a means to improve attribute reliability. ***Computers, Environment and Urban Systems*** (to appear)