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”)
Alpine
$52,917
+/- $13,603
= $39,314
– $66,520
APPROACHES HANDLING/INCORPORATING ERROR IN MAPPING

- Clearly acknowledging unreliability of data
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 American Community Survey
INDIVIDUAL FUNCTIONS OF THE ARCGIS EXTENSION

ACSD Mapping
- Download ACS Data and Shapefiles
- Join ACS Table(s) with Shapefiles
- Mapping Data Uncertainty
  - Overlay CVs with Estimates
    - Binary Mapping for any variable
    - Identify Areas of Significant Differences (from a selected estimate)
    - Identify Areas of Significant Differences (from all selected estimates)
    - Identify Areas of Significant Differences (from a fixed value)
    - Identify Areas of Significant Differences (from estimate(s) in a separate layer)
    - Compare Two Layers
- Documentation
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

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

6 The Washington Post Magazine • October 25, 2015
HANDLING/INCORPORATING ERROR IN MAPPING

- Clearly acknowledging unreliability
- Help map readers to discern if estimates are 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?
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

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

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

\[ S_{A,B} = \min_{i \in A, j \in B} (CL_{i,j}), \quad i \neq j, \]

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: \(\sim\) likelihood that estimates on 2 sides of a break value are different
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
IMPLEMENTATIONS BASED ON THE CLASS SEPARABILITY CRITERION


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
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.
RESULTANT MAP BY ADJUSTING CLASS BREAKS

<table>
<thead>
<tr>
<th>Separability</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.3%</td>
<td>275 - 368</td>
</tr>
<tr>
<td>80.6%</td>
<td>370 - 536</td>
</tr>
<tr>
<td>80.1%</td>
<td>538 - 796</td>
</tr>
<tr>
<td>796 - 4,014</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Separability</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.3%</td>
<td>275 - 368</td>
</tr>
<tr>
<td>80.1%</td>
<td>370 - 796</td>
</tr>
<tr>
<td>796 - 1,024</td>
<td></td>
</tr>
<tr>
<td>1,025 - 1,916</td>
<td></td>
</tr>
<tr>
<td>1,917 - 4,014</td>
<td></td>
</tr>
</tbody>
</table>
INCORPORATING ERROR IN MAPPING

Clearly acknowledging unreliability of data

- In choropleth maps, indicating how areal units are different between classes
- In choropleth maps, using classification methods that maximize the differences between classes after considering errors in estimates

- 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
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
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:


