American Community Survey (ACS) Toolbox

Elaine Hallisey, MA
Geographer / GIS Analyst
CDC/ATSDR

Esri Health GIS Conference
August 21, 2012
Background

- At the Centers for Disease Control and Prevention (CDC) we regularly use US Census data in many analyses relating to population health and safety.

- Sampling error resulting from the change to the American Community Survey (ACS) from the census long form (SF3), presents difficulties.

- ACS data users cannot ignore the error if they want statistically valid analysis results.
We provide a geoprocessing tool to help deal with sampling error issues.
The ACS Toolbox

- A geoprocessing tool that will:
  - Calculate margins of error (MOE) for user-derived estimates.
  - Calculate coefficient of variation (CV), a relative measure of sampling error.
  - Determine statistical difference among enumeration units over time or space.
  - Help determine a suitable classification scheme, for choropleth or other mapping, based on data uncertainty.
A margin of error (MOE) is provided for each ACS estimate.

The ACS MOE describes the precision of the estimate at the 90% confidence level (i.e. a 10% chance of an incorrect inference), the Census standard for published data.

For example, if the estimated number of mobile homes for a tract is 100 with a MOE of 67, then we can be 90% certain the tract has between 33 (100-67) and 167 (100+67) mobile homes.

This range, e.g. 33 to 167, is known as the confidence interval.
MOE Calculator

- Calculating MOEs can be fairly complex. The tool will calculate MOEs for:
  - Aggregated count data, i.e. two or more fields for individual enumerations units in the same feature layer that have been added to or subtracted from one another.
  - Derived proportions. The numerator of a proportion is a subset of the denominator. Example - The number of people in poverty divided by the total population.
  - Derived ratios. The numerator of a ratio is not a subset of the denominator. Example - The number of males with a college degree divided by the number of females with a college degree.
  - Confidence levels of 90, 95, or 99%.
The estimate, calculated by the user,
\[ MF5SUE = MU5E + M5E + FU5E + F5E \]
The output MOE, \( MF5UM = \sqrt{MU5M^2 + M5M^2 + FU5M^2 + F5M^2} \)
Relative Sampling Error – CV Calculator

- A coefficient of variation (CV) provides the relative amount of sampling error associated with a sample estimate. A CV is usually expressed as a percent.

- Because they are relative, CVs can be compared to one another.

- The lower the CV, the better. The National Research Council suggests a CV no higher than 12. Esri uses reliability threshold ranges of high (CV <=12%), medium (CV from 12 to 40%), and low (CV > 40%).

- The CV is a function of the overall sample size and the size of the population of interest.

- Smaller geographic units have higher sampling error. Multiyear estimates improve statistical reliability, i.e. they lower CVs.
## Relative Sampling Error – CV Calculator

### Estimate Reliability

- **High**
- **Medium**
- **Low**

### Output CV field

<table>
<thead>
<tr>
<th>FID</th>
<th>Shape</th>
<th>FIPS_1</th>
<th>Geography</th>
<th>MF5UE</th>
<th>MF5SUM</th>
<th>SE</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>152</td>
<td>Polygon</td>
<td>13001</td>
<td>Appling County, Georgia</td>
<td>680</td>
<td>214.841802</td>
<td>130.602919</td>
<td>19</td>
</tr>
<tr>
<td>147</td>
<td>Polygon</td>
<td>13003</td>
<td>Atkinson County, Georgia</td>
<td>401</td>
<td>96.239285</td>
<td>58.504125</td>
<td>15</td>
</tr>
<tr>
<td>154</td>
<td>Polygon</td>
<td>13005</td>
<td>Bacon County, Georgia</td>
<td>215</td>
<td>183.831445</td>
<td>111.751638</td>
<td>52</td>
</tr>
<tr>
<td>114</td>
<td>Polygon</td>
<td>13007</td>
<td>Baker County, Georgia</td>
<td>302</td>
<td>216.138844</td>
<td>131.391395</td>
<td>44</td>
</tr>
<tr>
<td>98</td>
<td>Polygon</td>
<td>13009</td>
<td>Baldwin County, Georgia</td>
<td>1168</td>
<td>290.442765</td>
<td>176.560951</td>
<td>15</td>
</tr>
</tbody>
</table>
Statistical Difference – Geo
Statistical Difference - Time

- Geo provides a critical value the user reviews to determine if a specified feature is significantly different from that variable for other features in a geographic area.

- Time provides a critical value the user reviews to determine if the values of user-specified variables over two different time periods, of the same length, are significantly different from one another. Although it is better to use non-overlapping time periods for multi-year estimates, the tool can account for overlapping time periods.

- We use the tests the Census Bureau recommends for determining statistical difference.
Statistical Difference – Geo

Statistical Difference from Highlighted County

- No Difference
- Different at 90%
- Different at 95%
- Different at 99%

Output Zdiff field

<table>
<thead>
<tr>
<th>FIPS</th>
<th>Geography</th>
<th>MF5UE</th>
<th>MF5UM</th>
<th>Orig_OID</th>
<th>CompareOID</th>
<th>ZDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>13077</td>
<td>Coweta County, Georgia</td>
<td>2004</td>
<td>304.047694</td>
<td>11</td>
<td>12</td>
<td>8.629755</td>
</tr>
<tr>
<td>13149</td>
<td>Heard County, Georgia</td>
<td>282</td>
<td>123.697211</td>
<td>12</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>13199</td>
<td>Meriwether County, Georgia</td>
<td>645</td>
<td>188.634567</td>
<td>13</td>
<td>12</td>
<td>2.64717</td>
</tr>
<tr>
<td>13285</td>
<td>Troup County, Georgia</td>
<td>1890</td>
<td>311.860546</td>
<td>14</td>
<td>12</td>
<td>7.884311</td>
</tr>
<tr>
<td>13231</td>
<td>Pike County, Georgia</td>
<td>196</td>
<td>58.017239</td>
<td>15</td>
<td>12</td>
<td>1.035445</td>
</tr>
</tbody>
</table>
Statistical Difference - Time

Input Feature Layer:
StateHis0709_0810_Overlap

Data Confidence Level:
90

Output Feature Class:
\cdc\project\VATS_GIS_Store4\Projects\prj04008_ACSTool\Docu

Number of Overlap Years:
2

--- First Time Period Fields ---
1a: Input Estimate Field
HISP1_E
1b: Input MOE Field
HISP1_M
1c: First time period year
2007
1d: Last time period year
2009
1e: Time Period
3 Years

--- Second Time Period Fields ---
2a: Input Estimate Field
HISP2_E
2b: Input MOE Field
HISP2_M
2c: First time period year
2008
2d: Last time period year
2010
2e: Time Period
3 Years

Output Zdiff field

Statistical Difference over Two Time Periods
- No MOE Provided
- No Difference
- Different at 90%
- Different at 95%
- Different at 99%

Table:

<table>
<thead>
<tr>
<th>FID</th>
<th>Shape</th>
<th>STATE_NAME</th>
<th>STATE_FIPS</th>
<th>HISP1_E</th>
<th>HISP1_M</th>
<th>HISP2_E</th>
<th>HISP2_M</th>
<th>ZDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Polygon</td>
<td>Florida</td>
<td>12</td>
<td>3882998</td>
<td>1319</td>
<td>4128985</td>
<td>586</td>
<td>485.598889</td>
</tr>
<tr>
<td>10</td>
<td>Polygon</td>
<td>Georgia</td>
<td>13</td>
<td>777579</td>
<td>1500</td>
<td>825908</td>
<td>700</td>
<td>83.187689</td>
</tr>
<tr>
<td>11</td>
<td>Polygon</td>
<td>Hawaii</td>
<td>15</td>
<td>113255</td>
<td>-999</td>
<td>118878</td>
<td>-999</td>
<td>-999</td>
</tr>
<tr>
<td>12</td>
<td>Polygon</td>
<td>Idaho</td>
<td>16</td>
<td>158341</td>
<td>98</td>
<td>169186</td>
<td>149</td>
<td>173.26402</td>
</tr>
</tbody>
</table>
**Classing Method Assistant**

- This tool assists the user in choosing a scheme based on data uncertainty from among natural breaks, quantiles, equal intervals, and manual classification.

- The tool’s algorithm is discussed in Konstantin Krivoruchko’s text *Spatial Statistical Data Analysis for GIS Users*. 
Selecting a suitable classification method for data with relatively large uncertainty can be difficult.
Classing Method Assistant

Geoprocessing results

Quantile Breaks: [0.0, 0.062741333333333302, 0.11099753333333333, 0.1844355, 1.0]
Class 1 Probability: 290.783786633
Class 2 Probability: 170.02839279
Class 3 Probability: 194.902839801
Class 4 Probability: 424.292203715
-->Total Probability for Quantile Breaks: 
1080.00722294

Equal Interval Breaks: ['0.0:0.25', '0.25:0.5', '0.5:0.75', '0.75:1.0']
Class 1 Probability: 1171.77161069
Class 2 Probability: 212.042209336
Class 3 Probability: 11.5655831477
Class 4 Probability: 1.40714067459
-->Total Probability for Equal Interval Breaks: 
1396.78654385

Natural Breaks(Jenks): ['0.0:0.10016', '0.10016:0.210953', '0.210953:0.362509', '0.362509:1.0']
Class 1 Probability: 537.521247275
Class 2 Probability: 343.745567185
Class 3 Probability: 193.279302922
Class 4 Probability: 87.7113786848
-->Total Probability for Natural Breaks(Jenks): 
1162.25749607
Project Team

Geospatial Research, Analysis & Services Program (GRASP) of CDC/ATSDR/DTHHS:

- **Jeff Henry & Andrew Chiang** – Developers
- **Brian Lewis, BS** – Statistician
- **Barry Flanagan, PhD** – Geographer
- **Marc Cunningham, MPH** (now at the John Snow Institute) – Research and Planning
- **Caitlin Mertzlufft, MPH** – Quality Control
- **Elaine Hallisey, MA** – Project Lead
References and recommended reading


References and recommended reading


References and recommended reading


References and recommended reading

Questions?

To obtain the ACS Toolbox, contact:
Elaine Hallisey
ehallisey@cdc.gov

Agency for Toxic Substances and Disease Registry

4770 Buford Highway NE, Chamblee, GA 30341
Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
E-mail: cdcinfo@cdc.gov Web: http://www.atsdr.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.