American Community Survey (ACS) Toolbox

Elaine Hallisey, MA
Geographer / GIS Analyst
CDC/ATSDR

Esri Health GIS Conference
August 21, 2012

Background

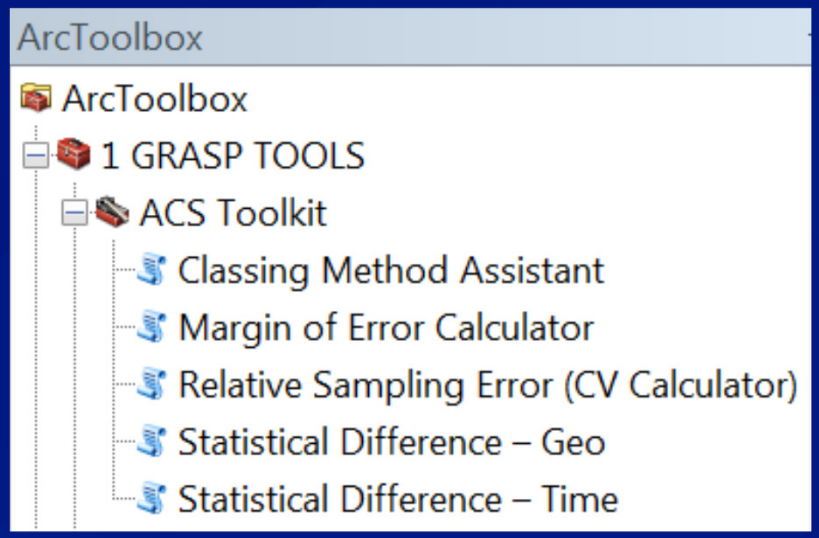
- At the Centers for Disease Control and Prevention (CDC) we regularly use US Census data in many analyses relating to population health and safety.
- Sampling error resulting from the change to the American Community Survey (ACS) from the census long form (SF3), presents difficulties.
- ACS data users cannot ignore the error if they want statistically valid analysis results.

Background

- 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.

MOE Calculator

- 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:**

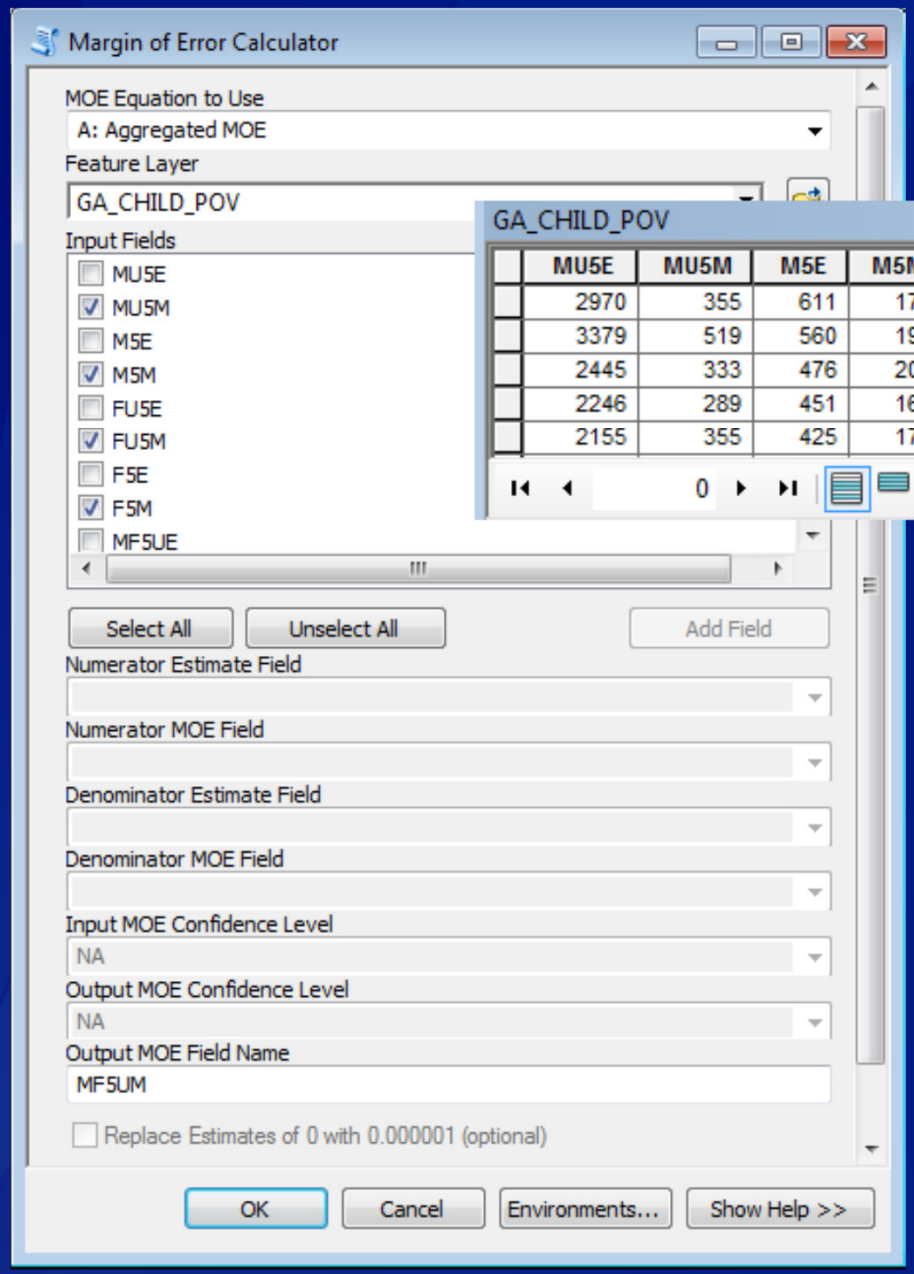
- Aggregated count data, i.e. two or more fields for individual enumerations units in the same feature layer that have been added to or subtracted from one another.

- 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%.

MOE Calculator



GA_CHILD_POV

MU5E	MU5M	M5E	M5M	FU5E	FU5M	F5E	F5M	MF5UE	MF5UM
2970	355	611	178	3144	384	603	181	7328	581
3379	519	560	195	2872	450	307	138	7118	727
2445	333	476	209	2642	443	406	172	5969	617
2246	289	451	165	1944	285	414	189	5055	477
2155	355	425	175	1860	303	289	128	4729	515

0 (0 out of 159 Selected)

Output MOE field

The estimate, calculated by the user,

MF5SUE =

MU5E + M5E + FU5E + F5E

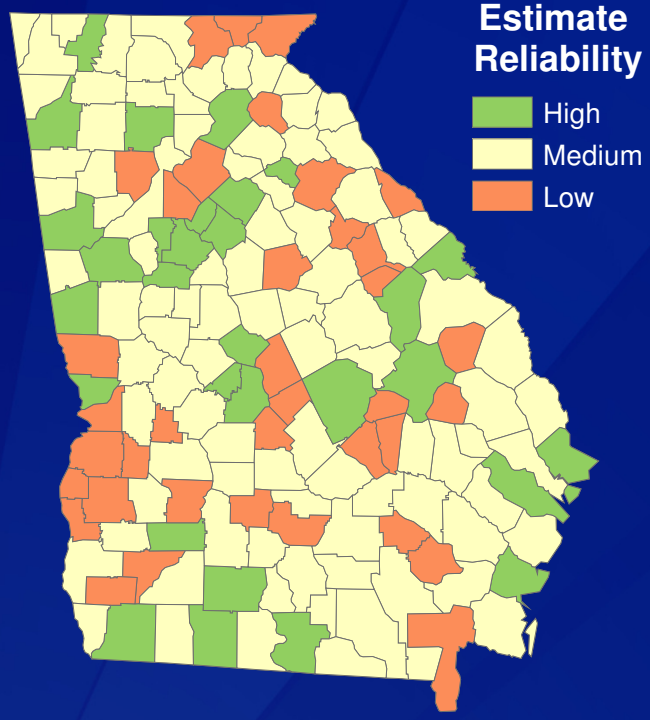
The output MOE, MF5UM =

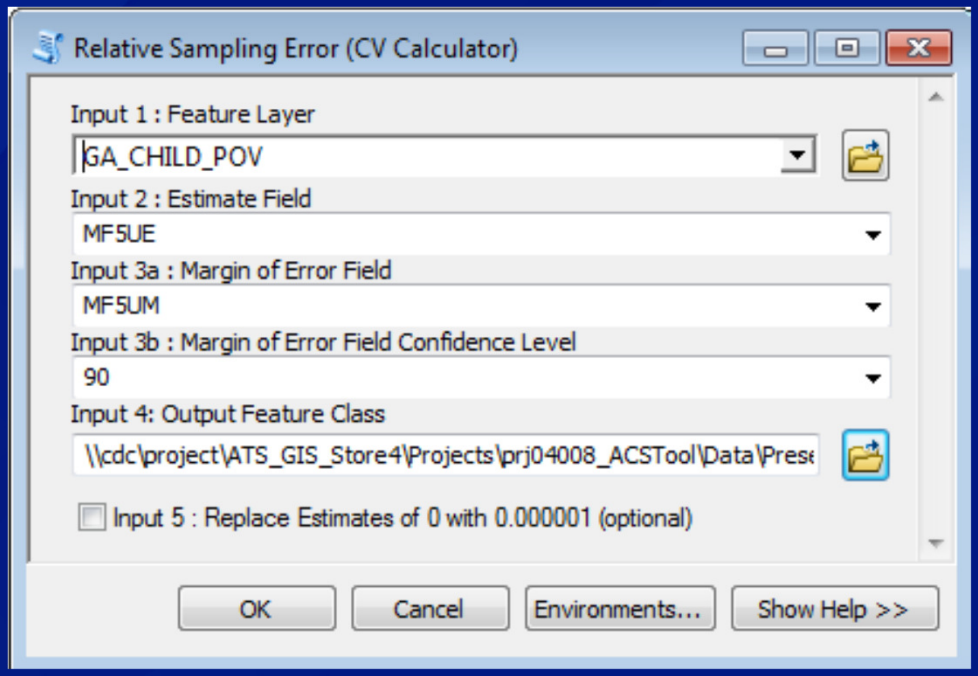
$\sqrt{MU5M^2 + M5M^2 + FU5M^2 + F5M^2}$

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





Output CV field

CVExample

	FID	Shape *	FIPS_1	Geography	MF5UE	MF5UM	SE	CV
	152	Polygon	13001	Appling County, Georgia	680	214.841802	130.602919	19
	147	Polygon	13003	Atkinson County, Georgia	401	96.239285	58.504125	15
	154	Polygon	13005	Bacon County, Georgia	215	183.831445	111.751638	52
	114	Polygon	13007	Baker County, Georgia	302	216.138844	131.391395	44
	98	Polygon	13009	Baldwin County, Georgia	1168	290.442765	176.560951	15

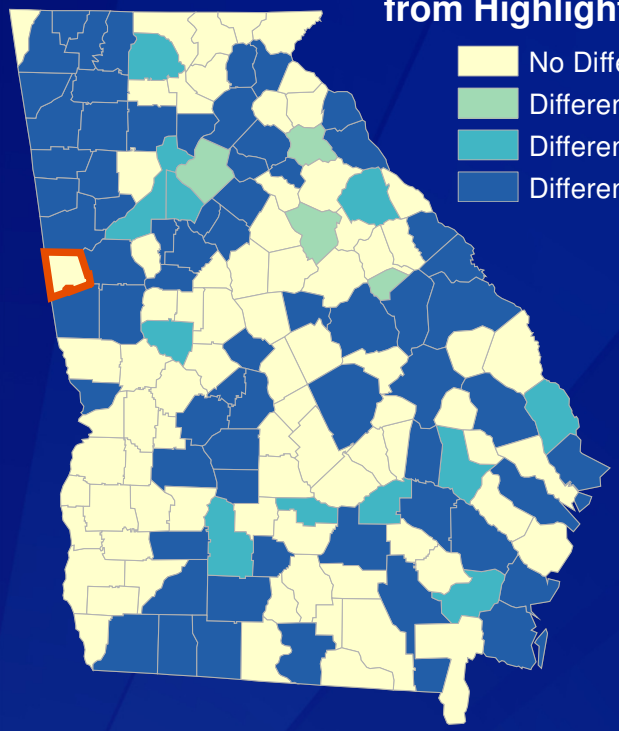
Statistical Difference – Geo
Statistical Difference - Time

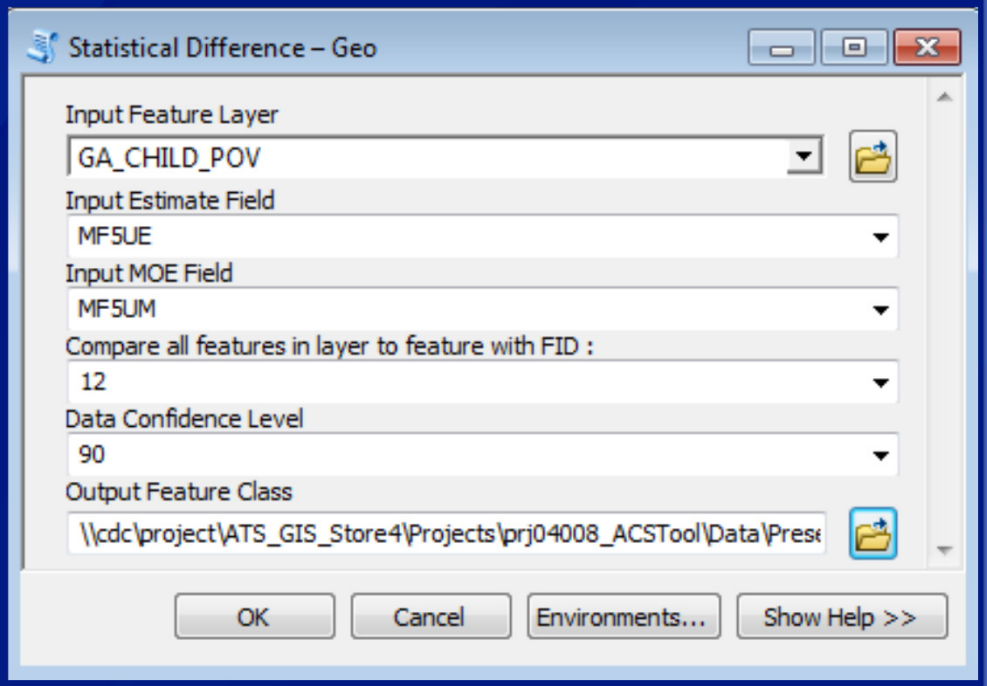
- Geo provides a critical value the user reviews to determine if a specified feature is significantly different from that variable for other features in a geographic area.

- 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



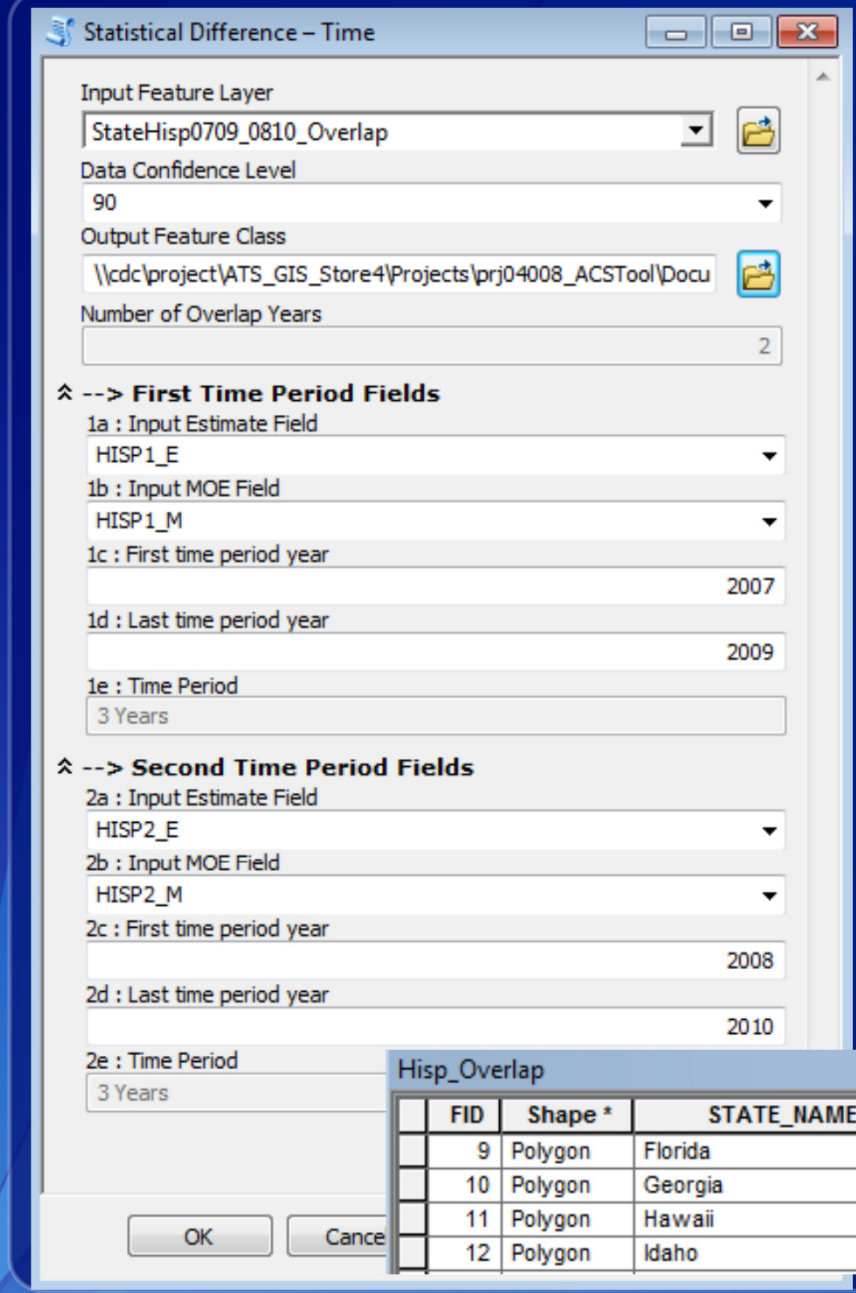


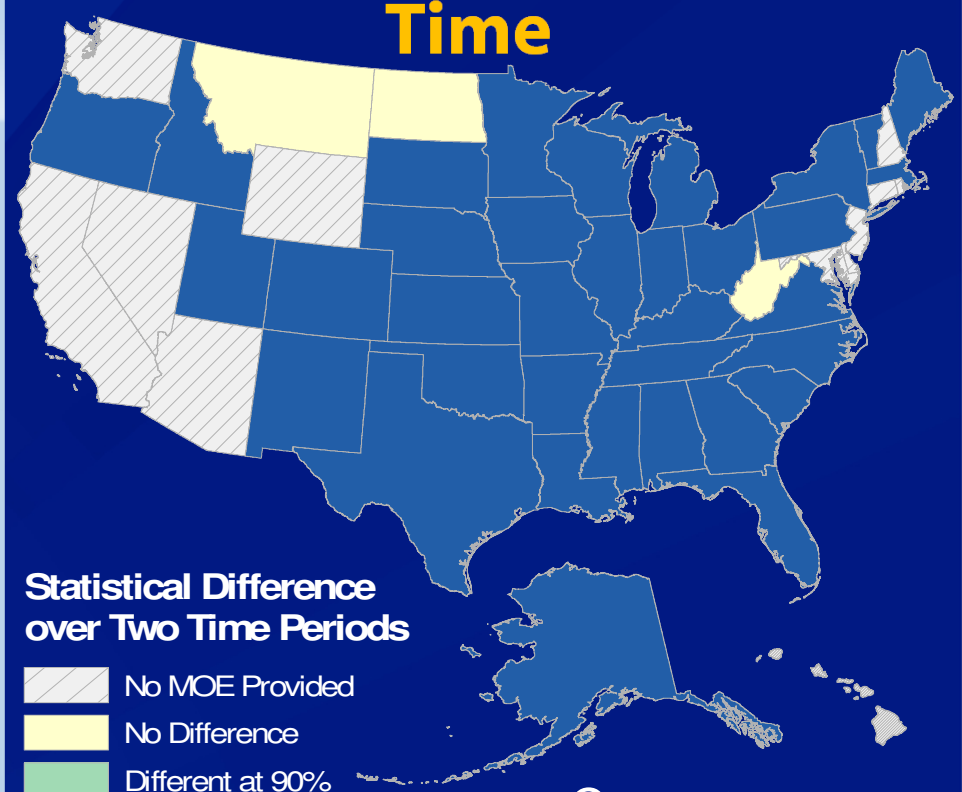
Output
Zdiff field

GeoExample

FIPS	Geography	MF5UE	MF5UM	Orig_OID	CompareOID	ZDiff
13077	Coweta County, Georgia	2004	304.047694	11	12	8.629755
13149	Heard County, Georgia	282	123.697211	12	12	0
13199	Meriwether County, Georgia	645	188.634567	13	12	2.64717
13285	Troup County, Georgia	1890	311.860546	14	12	7.884311
13231	Pike County, Georgia	196	58.017239	15	12	1.035445

Statistical Difference - Time





Output Zdiff field

Hisp_Overlap

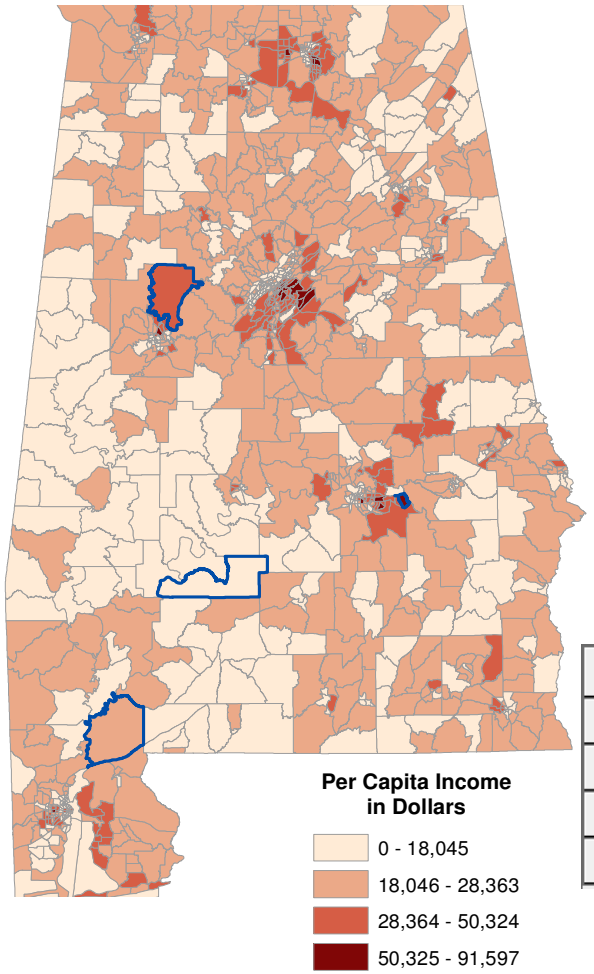
	FID	Shape *	STATE_NAME	STATE_FIPS	HISP1_E	HISP1_M	HISP2_E	HISP2_M	ZDiff
	9	Polygon	Florida	12	3882998	1319	4128985	586	485.598889
	10	Polygon	Georgia	13	777579	1500	825908	700	83.187689
	11	Polygon	Hawaii	15	113255	-999	118878	-999	-999
	12	Polygon	Idaho	16	158341	98	169186	149	173.26402

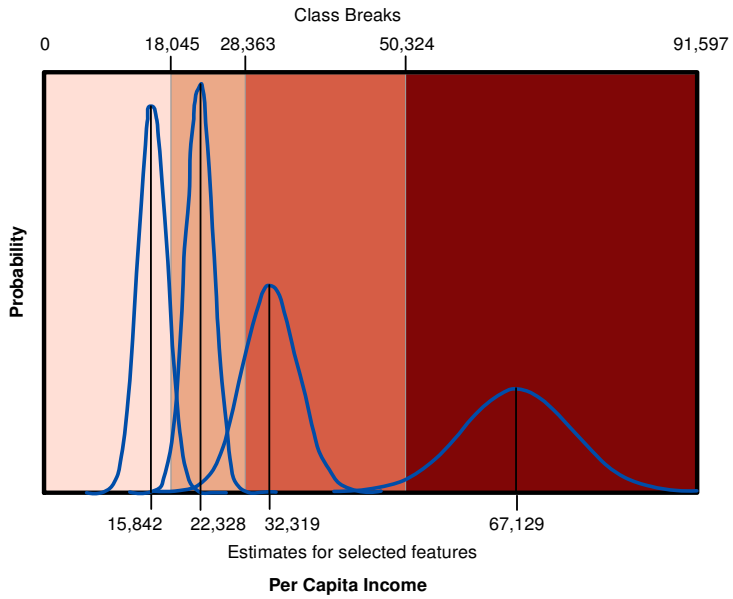
Classing Method Assistant

- 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*.

Classing Method Assistant



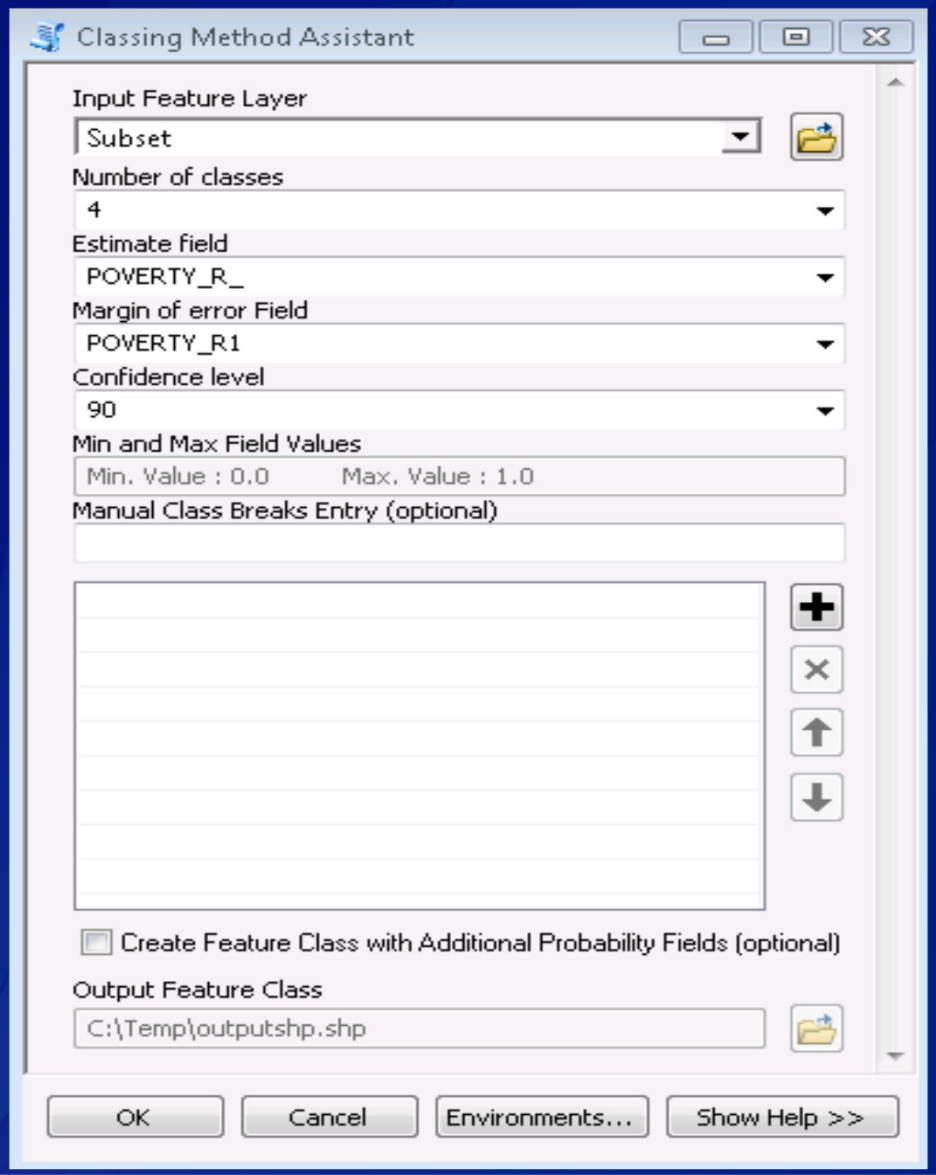


Selected Features

	GEOID	Estimate	MOE_99	Lower	Upper
	01131	15,842	5,174	10,668	21,016
	01003	22,328	4,916	17,412	27,244
	01125	32,319	9,930	22,389	42,249
	01101	67,129	19,823	47,306	86,952

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

- Esri White Paper (2011). *The American Community Survey*. http://www.esri.com/library/whitepapers/pdfs/the-american-community-survey.pdf

- Krivoruchko, Konstantin. 2011. Spatial Statistical Data Analysis for GIS Users. Esri Press.

- MacDonald, Heather. 2006. The American Community Survey: Warmer (More Current), but Fuzzier (Less Precise) than the Decennial Census. Journal of the American Planning Association, Vol. 72, No. 4.

- National Research Council. (2007). Using the American Community Survey: Benefits and Challenges. http://www.nap.edu/catalog.php?record_id=11901

References and recommended reading

- Sun, Min and D. Wong. 2010. Incorporating Data Quality Information in Mapping American Community Survey Data. Cartography and Geographic Information Science, Vol. 37, No. 4, 2010, pp. 285-300.

- U.S. Census Bureau, *A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know*, http://www.census.gov/acs/www/Downloads/handbooks/ACSGeneralHandbook.pdf

- U.S. Census Bureau, *Things That May Affect Estimates from the American Community Survey*, http://www.census.gov/acs/www/Downloads/presentations/ACS_Affect_Est.ppt

References and recommended reading

- U.S. Census Bureau, Instructions for Applying Statistical Testing to the 2008-2010 3-Year Data and the 2006-2010 ACS 5-Year Data, http://www.census.gov/acs/www/Downloads/data_documentation/Statistical_Testing/2010StatisticalTesting3and5year.pdf

- Wombold, Lynn. Esri. (2007). Changes and Challenges: *Understanding the American Community Survey*. ArcUser. http://www.esri.com/news/arcuser/1207/census.html

- Wombold, Lynn. Esri. (2008). *Sample Size Matters: Caveats for users of ACS tabulations.* ArcUser. http://www.esri.com/library/reprints/pdfs/arcuser_sample-size.pdf

References and recommended reading

- Xiao, Ningchuan, C. Calder, and M. Armstrong. 2007. Assessing the effect of attribute uncertainty on the robustness of choropleth map classification. International Journal of Geographical Information Science. Vol. 21, No. 2, February 2007, 121–144.

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