Alternative Strategies for Mapping ACS Estimates and Error of Estimation

Joe Francis, Jan Vink, Nij Tontisirin, Sutee Anantsuksomsri & Viktor Zhong

Cornell University
Program on Applied Demographics

Acknowledgements

- Thanks for all the work on geovisualization ideas and map design at PAD shown here:
  - Jan Vink,
  - Nij Tontisirin,
  - Sutee Anantsuksomsri,
  - Viktor Zhong
- Appreciation of support from the
  - Cornell Population Center, especially the director, Dan Lichter
  - Dept. of Development Sociology, especially the chair, David Brown

Introduction

- ACS now THE primary mechanism
  - for measuring and disseminating detailed socio-economic characteristics of the population at the sub-state level
  - smaller geographies like tracts
  - In it's second iteration of 1,3,5 year releases
  - Sampling and measurement source of error less.

Introduction

- With ACS, the Census Bureau began to report forthrightly both
  - the estimates
  - uncertainty of their sample estimates
- Presenting both components to an audience is a challenge.
- Many people don’t report error or bury it in appendix.
- Not good practice.

Introduction

- Some recent work on presenting error levels along with estimates in spreadsheets.
- Uses classification and color coding.
- Here’s a couple of ideas of how to present both for spreadsheets.
- The first is from ESRI
- The second is from a Census Bureau Usability research on ACS.

Introduction

ESRI's reliability symbols are as follows:

High Reliability: The ACS estimate is considered to be reliable. The sampling error is small relative to the estimate.

Medium Reliability: Use the ACS estimate with caution. The sampling error is fairly large relative to the estimate.

Low Reliability: The ACS estimate is considered unreliable. The sampling error is very large relative to the estimate.

In

- Example of an ESRI embellished spreadsheet.

esri

ACS Population Summary

Humboldt County, CA
Humboldt County, CA (06023)
Geography: County

	2005 - 2009 ACS Estimate	Pe	MOE(±)	Reliability
POPULATION AGE 5+ YEARS BY LANGUAGE SPOKEN AT HOME AND ABILITY TO SPEAK ENGLISH				
Total	121,393		41	
5 to 17 years				
Speak only English	16,914		346	
Speak Spanish	1,400		254	
Speak English "very well" or "well"	1,309		278	
Speak English "not well"	83		64	
Speak English "not at all"	8	0.	12	
Speak other Indo-European languages	207	0.2%		
Speak English "very well" or "well"	184	0.2%		
Speak English "not well"	23	0.0%		
Speak English "not at all"	0	0.0%	132	
Speak Asian and Pacific Island languages	240	0.2%	111	
Speak English "very well" or "well"	228	0.2%	110	
Speak English "not well"	12	0.0%	19	
Speak English "not at all"	0	0.0%	132	
Speak other languages	126	0.1%	95	
Speak English "very well" or "well"	109	0.1%	90	
Speak English "not well"	17	0.0%	28	
Speak English "not at all"	0	0.0%	132	
18 to 64 years				
Speak only English	78,751	64.9%	683	
Speak Spanish	4,488	3.7%	490	
Speak English "very well" or "well"	3,068	2.5%	372	
Speak English "not well"	984	0.8%	275	
Speak English "not at all"	436	0.4%	160	
Speak other Indo-European languages	1,307	1.1%	323	
Speak English "very well" or "well"	1,056	0.9%	253	
Speak English "not well"	190	0.2%	117	
Speak English "not at all"	61	0.1%	64	
Speak Asian and Pacific Island languages	1,175	1.0%	233	
Speak English "very well" or "well"	948	0.8%	227	
Speak English "not well"	129	0.1%	76	
Speak English "not at all"	98	0.1%	87	
Speak other languages	463	0.4%	234	
Speak English "very well" or "well"	463	0.4%	236	
Speak English "not well"	0	0.0%	132	
Speak English "not at all"	0	0.0%	132	
65 years and over				
Speak only English	15,071	12.4%	192	
Speak Spanish	380	0.3%	73	
Speak English "very well" or "well"	250	0.2%	87	
Speak English "not well"	78	0.1%	53	
Speak English "not at all"	52	0.0%	49	
Speak other Indo-European languages	674	0.6%	166	
Speak English "very well" or "well"	587	0.5%	158	
Speak English "not well"	58	0.0%	45	
Speak English "not at all"	29	0.0%	43	
Speak Asian and Pacific Island languages	141	0.1%	56	
Speak English "very well" or "well"	73	0.1%	37	
Speak English "not well"	30	0.0%	40	
Speak English "not at all"	38	0.0%	45	
Speak other languages	56	0.0%	37	
Speak English "very well" or "well"	56	0.0%	137	
Speak English "not well"	0	0.0%	132	
Speak English "not at all"	0	0.0%	132	

Source: U.S. Census Bureau, 2005-2009 American Community Survey

Reliability: high medium low

March 23, 2011

Made with Esri Business Analyst

www.esri.com/ba 800-447-9778 Try it Now!

Two-Level Indicator Tables

Selected Social Characteristics in the United States: 2006
Data Set: 2006 American Community Survey
Survey: 2006 American Community Survey
Geographic Area: New York city, New York

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

NOTE ABOUT RELIABILITY OF SOME ESTIMATES:

Reliability
use caution

Estimates marked with a yellow "use caution" reliability box (shown above) means that the coefficient of variation (CV) is above 0.30. The CV is defined as the standard error of an estimate divided by the mean of that estimate, measured as a percentage. Relatively, a lower CV means a more reliable estimate.

Selected Social Characteristics in the United States: 2006	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	**3,020,284**		**+/-9,874**
Family households (families)	1,823,095		+/-15,027
With own children under 18 years	853,419		+/-12,447
Married-couple families	1,089,597		+/-13,850
With own children under 18 years	502,221		+/-11,080
Male householder, no wife present	164,869		+/-6,286
With own children under 18 years	53,881		+/-4,642
Female householder, no husband present	568,629		+/-11,235
With own children under 18 years	297,317		+/-9,022
Nonfamily households	1,197,189		+/-13,299
Householder living alone	1,005,277		+/-13,313
65 years and over	310,098		+/-7,156
Households with one or more people under 18 years	975,699		+/-13,037
Households with one or more people 65 years and over	727,931		+/-5,676
Average household size	2.66		+/-0.01
Average family size	3.48		+/-0.02
RELATIONSHIP			
Household population	**8,035,586**		**+/-1,713**
Householder	3,020,284		+/-9,874
Spouse	1,087,632		+/-13,872
Child	2,528,284		+/-18,412
Other relatives	906,358		+/-15,851
Nonrelatives	493,028		+/-13,850
Unmarried partner	142,518		+/-6,160
MARITAL STATUS			
Males 15 years and over	**3,085,630**		**+/-790**
Never married	1,340,477		+/-12,623
Now married, except separated	1,388,250		+/-14,189
Separated	90,327		+/-5,221
Widowed	74,161		+/-4,027
Divorced	192,415		+/-6,306

Three-Level Indicator Tables

Selected Social Characteristics in the United States: 2006
Data Set: 2006 American Community Survey
Survey: 2006 American Community Survey
Geographic Area: Cook County, Illinois

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Selected Social Characteristics in the United States: 2006	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	**1,932,197**	good	**+/-8,605**
Family households (families)	1,217,733	good	+/-11,714
With own children under 18 years	576,417	good	+/-9,573
Married-couple families	824,922	good	+/-12,260
With own children under 18 years	380,049	good	+/-9,213
Male householder, no wife present	97,227	good	+/-4,702
With own children under 18 years	36,162	good	+/-2,913
Female householder, no husband present	295,584	good	+/-7,577
With own children under 18 years	160,206	good	+/-5,923
Nonfamily households	714,464	good	+/-10,203
Householder living alone	607,726	good	+/-10,372
65 years and over	183,241	good	+/-5,319
Households with one or more people under 18 years	650,125	good	+/-9,711
Households with one or more people 65 years and over	445,793	good	+/-3,588
Average household size	2.69	good	+/-0.01
Average family size	3.47	good	+/-0.03
RELATIONSHIP			
Household population	**5,197,637**	good	**+/-684**
Householder	1,932,197	good	+/-8,605
Spouse	823,694	good	+/-12,275
Child	1,685,734	good	+/-17,554
Other relatives	499,106	good	+/-17,284
Nonrelatives	256,906	good	+/-10,099
Unmarried partner	92,715	good	+/-5,128
MARITAL STATUS			
Males 15 years and over	**1,998,715**	good	**+/-386**
Never married	816,703	good	+/-10,439
Now married, except separated	941,158	good	+/-13,028
Separated	43,162	good	+/-3,652
Widowed	53,199	good	+/-3,758
Divorced	144,493	good	+/-5,678

Reliability Legend based on the Coefficient of Variation (CV)

Range	Reliability
CV <=0.30	good
0.30 < CV <=0.61	fair
CV > 0.61	poor

The coefficient of variation (CV) is defined as the standard error of an estimate divided by the mean of that estimate, measured as a percentage. Relatively, a lower CV means a more reliable estimate.

Four-Level Indicator Tables

Selected Social Characteristics in the United States: 2006
Data Set: 2006 American Community Survey
Survey: 2006 American Community Survey
Geographic Area: California

NOTE. Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

Reliability Legend based on the Coefficient of Variation (CV)

Range	Reliability
CV <=0.10	excellent
0.10 < CV <=0.30	good
0.30 < CV <=0.61	fair
CV > 0.61	poor

The coefficient of variation (CV) is defined as the standard error of an estimate divided by the mean of that estimate, measured as a percentage. Relatively, a lower CV means a more reliable estimate.

Selected Social Characteristics in the United States: 2006	Estimate	Reliability	Margin of Error
HOUSEHOLDS BY TYPE			
Total households	**12,151,227**	excellent	**+/-18,090**
Family households (families)	8,303,793	excellent	+/-26,257
With own children under 18 years	4,239,440	excellent	+/-24,678
Married-couple families	6,051,701	excellent	+/-25,589
With own children under 18 years	3,010,321	excellent	+/-21,578
Male householder, no wife present	698,432	excellent	+/-14,287
With own children under 18 years	338,125	excellent	+/-10,354
Female householder, no husband present	1,553,660	excellent	+/-16,324
With own children under 18 years	890,994	excellent	+/-13,084
Nonfamily households	3,847,434	excellent	+/-25,448
Householder living alone	2,994,372	excellent	+/-23,715
65 years and over	978,553	excellent	+/-11,756
Households with one or more people under 18 years	4,696,427	excellent	+/-27,592
Households with one or more people 65 years and over	2,710,892	excellent	+/-11,227
Average household size	2.93	excellent	+/-0.01
Average family size	3.54	excellent	+/-0.01
RELATIONSHIP			
Household population	**35,594,342**	excellent	*****
Householder	12,151,227	excellent	+/-18,090
Spouse	6,046,430	excellent	+/-26,006
Child	11,567,876	excellent	+/-38,674
Other relatives	3,452,551	excellent	+/-40,572
Nonrelatives	2,376,258	excellent	+/-33,162
Unmarried partner	743,837	excellent	+/-12,746
MARITAL STATUS			
Males 15 years and over	**14,185,501**	excellent	**+/-4,738**
Never married	5,374,190	excellent	+/-25,966
Now married, except separated	7,097,824	excellent	+/-29,939
Separated	280,535	excellent	+/-9,066
Widowed	308,020	excellent	+/-8,095
Divorced	1,124,932	excellent	+/-14,803

Introduction

- Here we focus on how to create maps that include information about the sampling error
- Currently the most prevalent practice is to largely ignore the unreliability of ACS estimates when mapping.
- Partially this is a result of the difficulty users have with interpreting maps.
- This needs to change if users of our maps are to place confidence in our map making.

Introduction

- Visualization of uncertainty data is a challenge we should not walk away from
- Begin by acknowledging that all survey and GIS data have error to some degree and there are many reasons for its presence.
- The question before us is not whether to present this information of uncertainty in our estimation but how.

Introduction

- GIS and cartographers have worked on the problem of how to present uncertainty of data values for over two decades.
- Kardos, Moore and Benwell (2003) have provided a nice summary of work that has been done.
- Not on that list is recent work by Stewart and Kennelly (2010) on use of 3-D “prisms” and “shadowing” to convey uncertainty.

Mode	Name	Description
Static	Adjacent Maps	Two maps can be used to show the uncertainty, one to show the actual information and another to show the uncertainty (MacEachren et al 1998)
	Overlay	A single choropleth map can be used to show the attribute information with an overlay of the uncertain information shown as textures on top (MacEachren 1992; MacEachren, et al 1998).
	Blurring	The clarity of an area boundary is used to define the uncertainty of the spatial data. A sharp pattern would indicate certain information; an approximate pattern definition would indicate uncertain information (MacEachren 1992).
	Fog	Uncertain parts on a map are partially hidden, therefore unclear to see. The thicker the fog the more uncertainty is in that part of the map. Fog obscures data from viewing; this is not an issue since such uncertain data is of no inherent use and thus should not be seen by the user (MacEachren 1992).
	Pixel Mixture	Pixels are divided into sub-pixels and an appropriate class value is given to each sub-pixel proportional to the membership function calculation (De Gruijter, et al 1997).
	Saturation of Colour	Saturation of colour is used to visualise uncertainty. The more saturated (richer) a colour representing a particular class, the more certain the information is on that part of the map. (Hengl, et al 2002).
	Sound	This provided a level of uncertainty at a particular location on a map through a variable pitch. A low pitch sound depicted low uncertainty and a high pitch sound for large uncertainty (cursor-driven) (Fisher 1994; Krygier 1994).
	Trustree	The outline of a quadtree structure is used to characterize the changing uncertainty across a choropleth boundary; the semi-continuous change across a quadtree, from cell to cell helps dilute the arbitrary assumptions associated with choropleth boundaries (Kardos et al, 2005)
Dynamic	Blinking pixels	Information in the spatial display was manipulated causing it to blink, hence highlighting those uncertain areas to the viewer (through more rapid blinking) (Fisher 1993; Monmonier & Gluck 1994; Evans 1997).
	Animation	A movie clip of map realisations (generated from a Monte Carlo simulation) highlighting areas where data is considered to be uncertain. Ehlschlaeger et al. (1997) state that if there is little change between realisations then one can be fairly convinced about the extent of the uncertainty.

Introduction

- These efforts have much to inform our present dilemmas.
- The work of Sun and Wong(2010) as well as Torrieri, Wong and Ratcliffe (2011) are examples of some recent attempts to deal with the geo-visualization problem.
- We think it would be a mistake to foreclose too quickly on one system for presenting ACS estimates and errors of estimation.
- We would like to present some alternative ideas.

Estimation Error in ACS

- 10 major issues to deal with in portraying estimation uncertainty in the ACS, SAIPE and similar sample survey data.

1. *Absolute vs. Relative Error*
2. *Side-by-Side maps vs. Overlay Maps*
3. *Crisp vs. Modified Classes*
4. *Number of Classes*
5. *Method of Classification*
6. *Symbolizing Uncertainty*

Estimation Error in ACS

7. *Map Legends*
8. *Static vs. dynamic interactive maps*
9. *Number of geographic units on map*
10. *Map Complexity and Type of User*

- Here we would like to offer a few comments on some of these issues.
- Our background paper contains more detailed comments on these and rest.

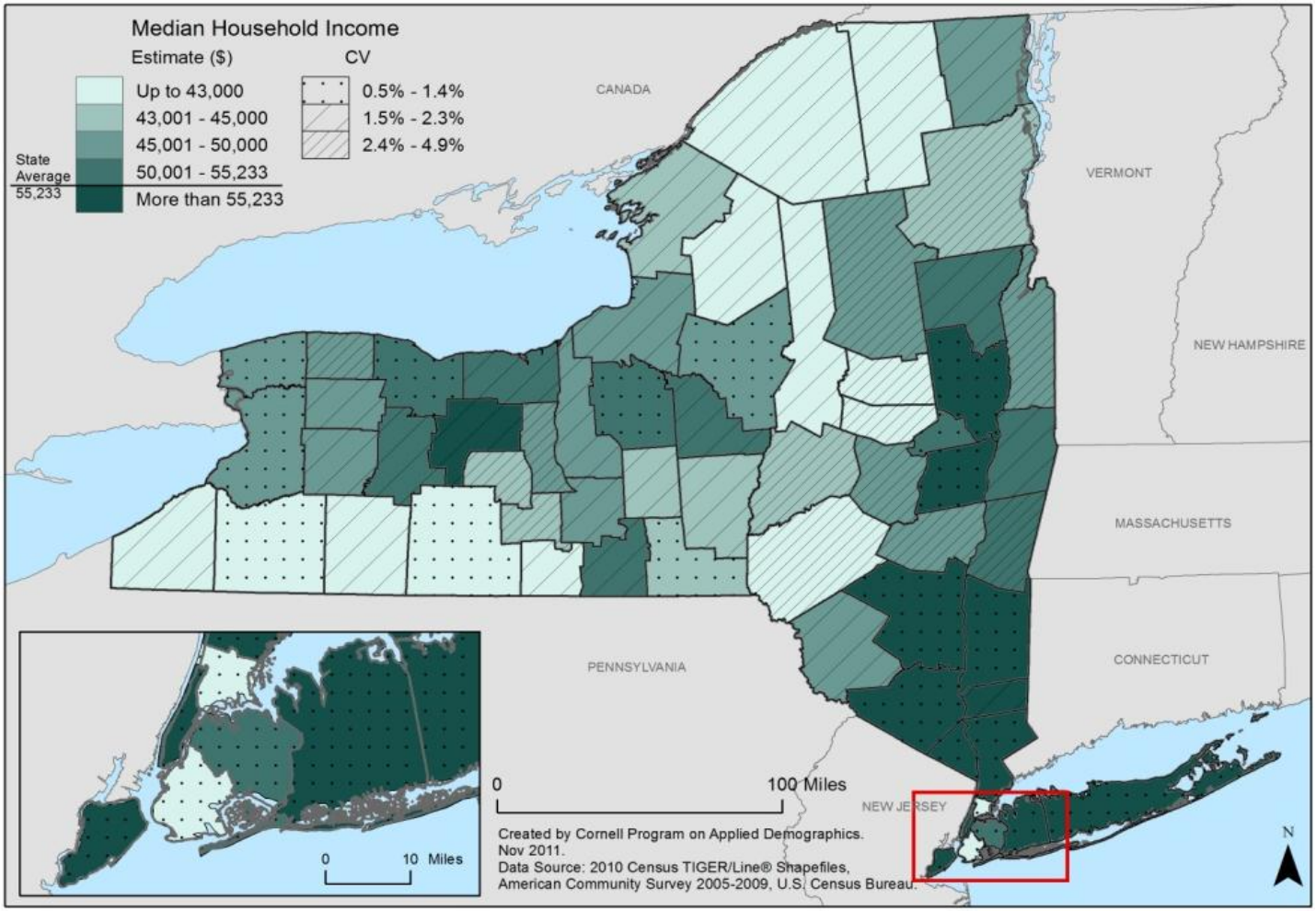
Absolute vs. Relative Error

- First issue: what to use as measure of error
- Some researchers argue for the use of relative error rather than absolute error measures.
- The reason—absolute error measures are sensitive to the scale of the estimate.
- Worry is that less careful user will focus only on the size of the error and draw conclusion that big error always signals high unreliability without taking into account the scale of the data or the estimate thereof.

Absolute vs. Relative Error

- While acknowledging that can be a problem, we feel that unmindful use of the CV has problems as well.
- Our work leads us to conclude that choice depends on the format of the variable being estimated.
- For totals, medians and mean averages, use of relative measures of error like the coefficient of variation (CV) seems more appropriate.

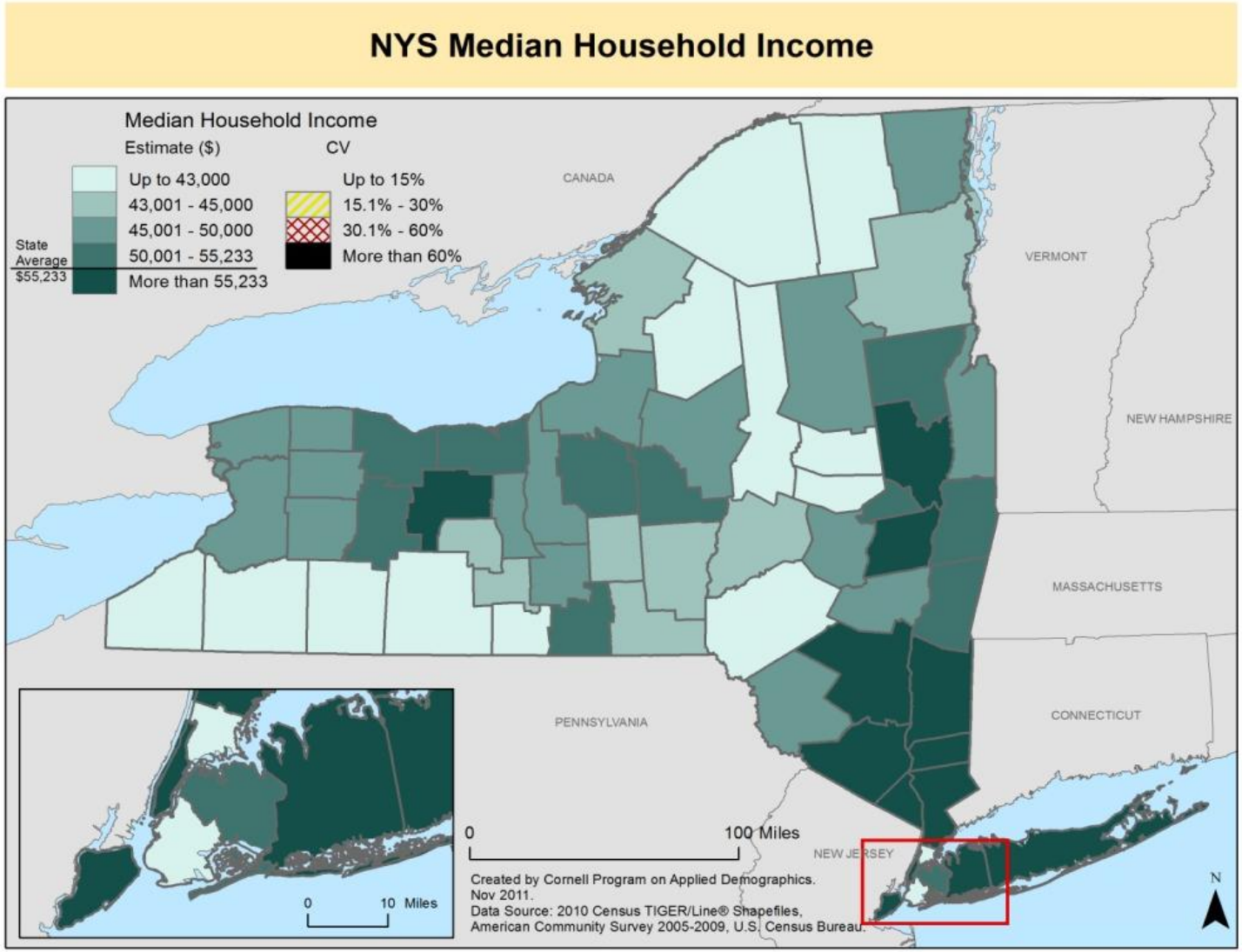
Number of Geographic Units



Absolute vs. Relative Error

- Relative error good for
  - measuring stability/reliability
  - comparison between types of data or with different dimensions
  - comparison between estimates of different orders of magnitude
  - if possible outcomes are
    - Bounded $[0, +\infty]$
    - Quantitative level measurement (not categorical)

Number of Geographic Units



Absolute vs. Relative Error

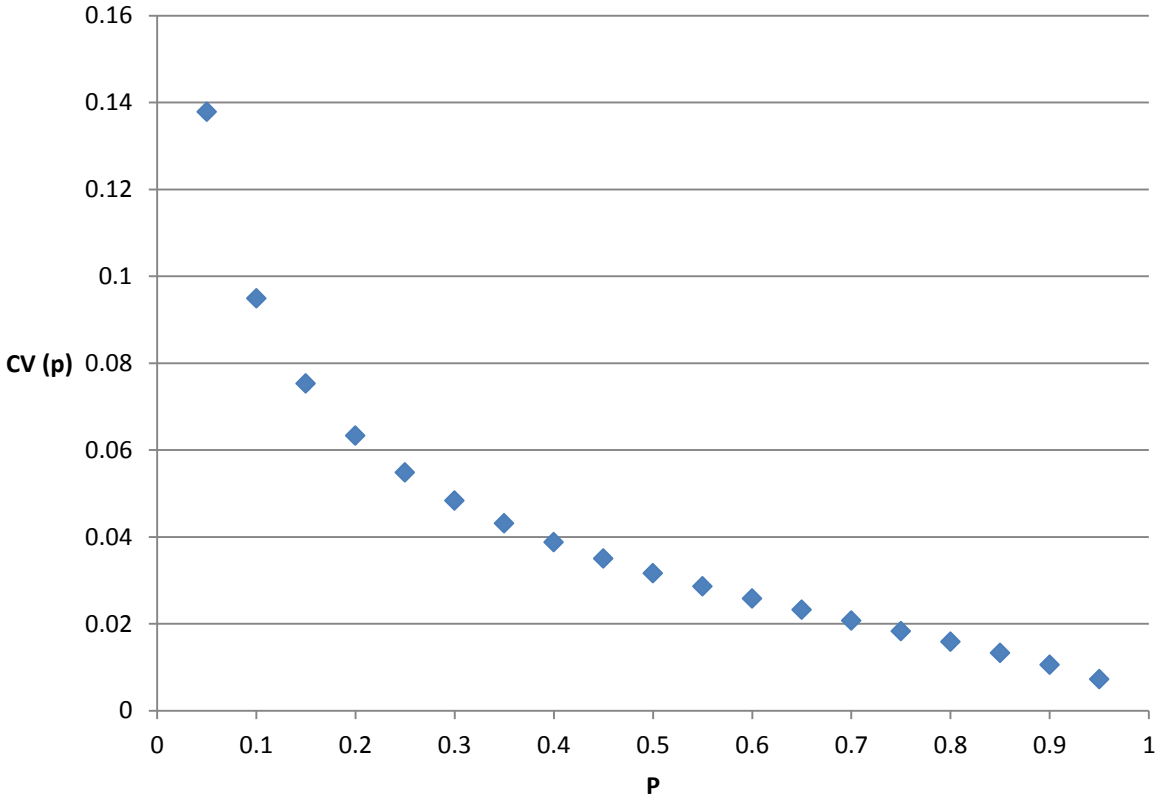
- However, for proportions, %'s, or a ratio like sex ratio, the standard confidence interval seems the more appropriate measure.
- Because proportions are bounded by 0 to 1, the CV presents interpretation problems.
  - Becomes unstable when estimate approaches 0, or when it approaches 1
  - Confusing for estimate with ranges [- ∞, +∞], like when estimating change over time

Absolute vs. Relative Error

- To illustrate, consider a variable like foreign born where one geographic unit has an estimate of 10% with MOE of ±8% and a second geography with an estimate of 90% with MOE of ±8%.
- Though structurally equivalent, the CV for the 10% foreign born is 48% (very unreliable) while the CV for the 90% native born is 5% (very reliable). Does this make sense?

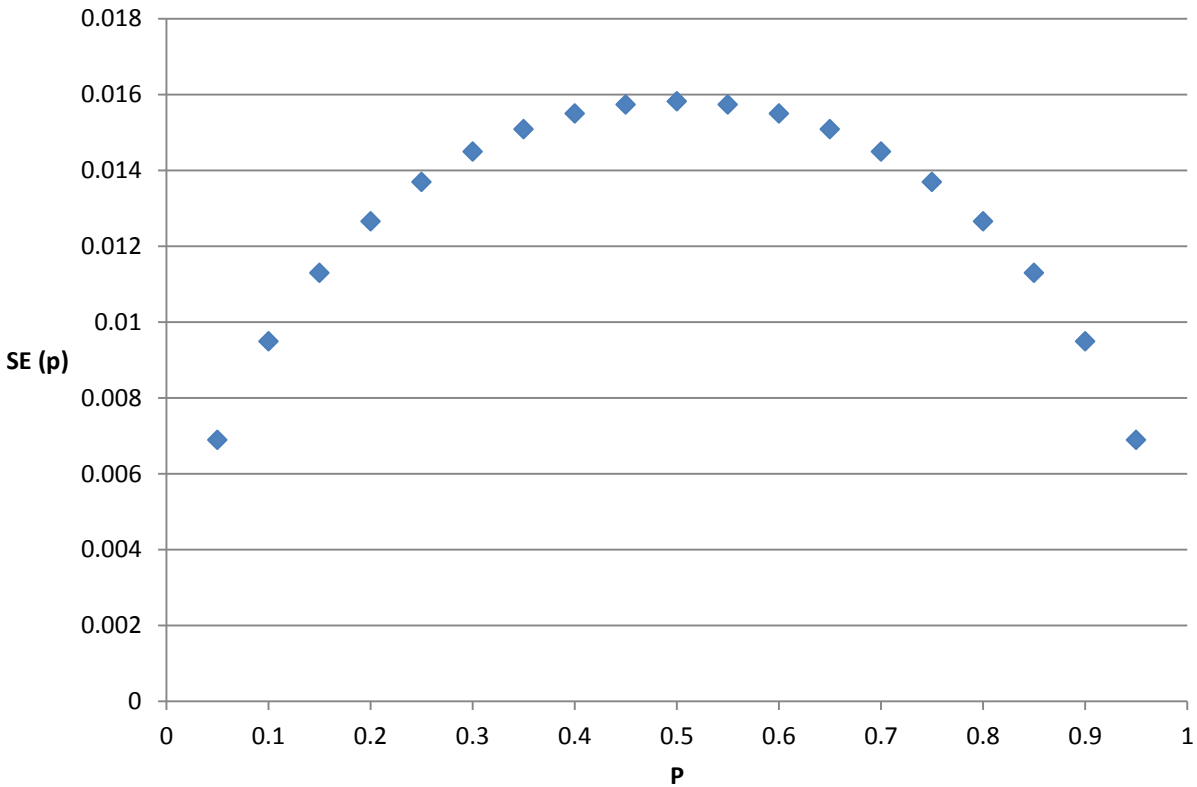
Absolute vs. Relative Error

- Part of the problem may lay in asymmetrical, nonlinear nature of the distribution of CV.



Absolute vs. Relative Error

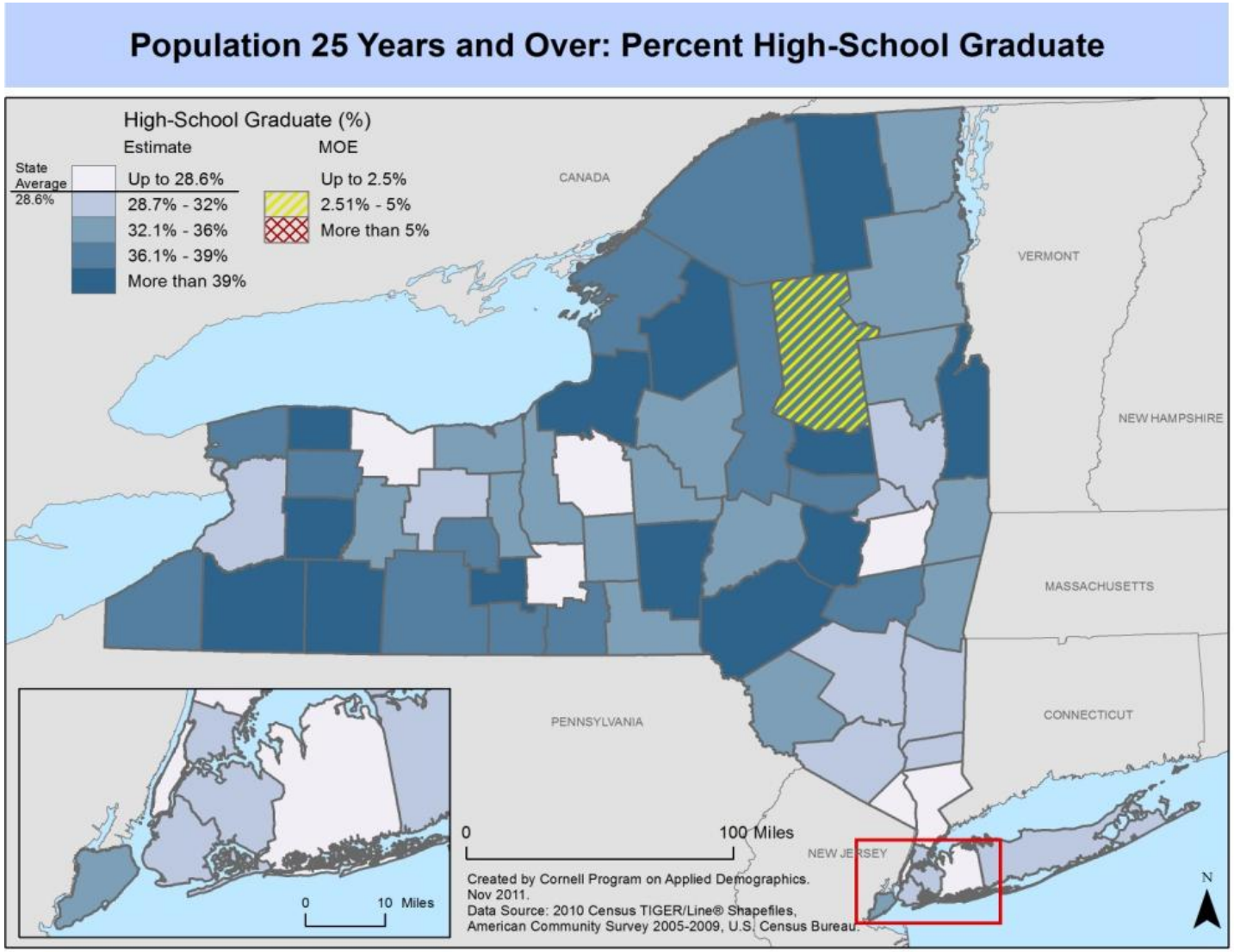
- On the other hand, for variables like this, the confidence interval performs as expected.



Absolute vs. Relative Error

- For both the estimate of p =10% foreign born and q=90% native born, the standard error of estimate is the same, approximately 0.01 when n=1000.
- This symmetry for placing a confidence bound on the estimate makes more sense both intuitively and statistically to us compared to a nonlinear relative error measure like the CV.
- So we choose to use the MOE in these circumstances, as illustrated next.

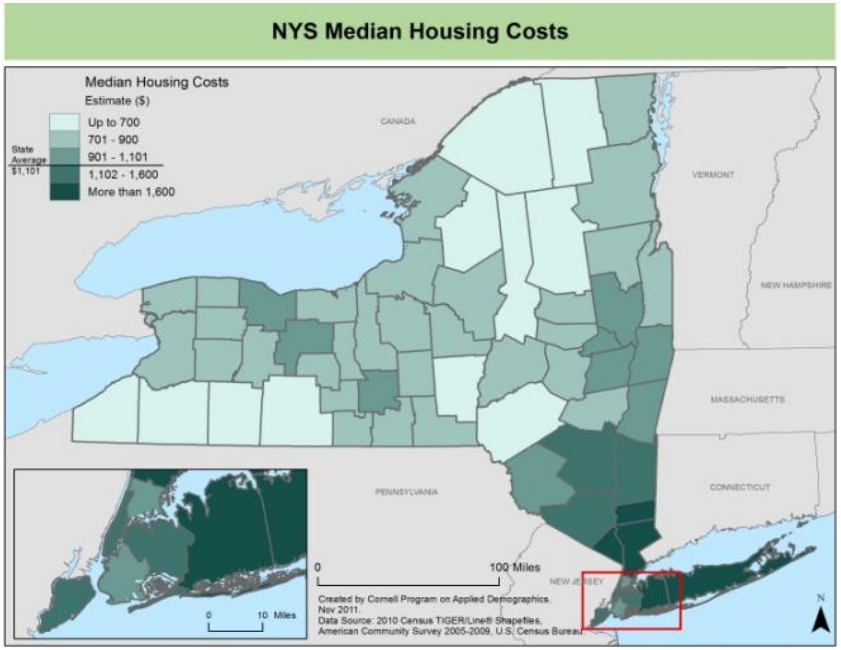
Absolute vs. Relative Error

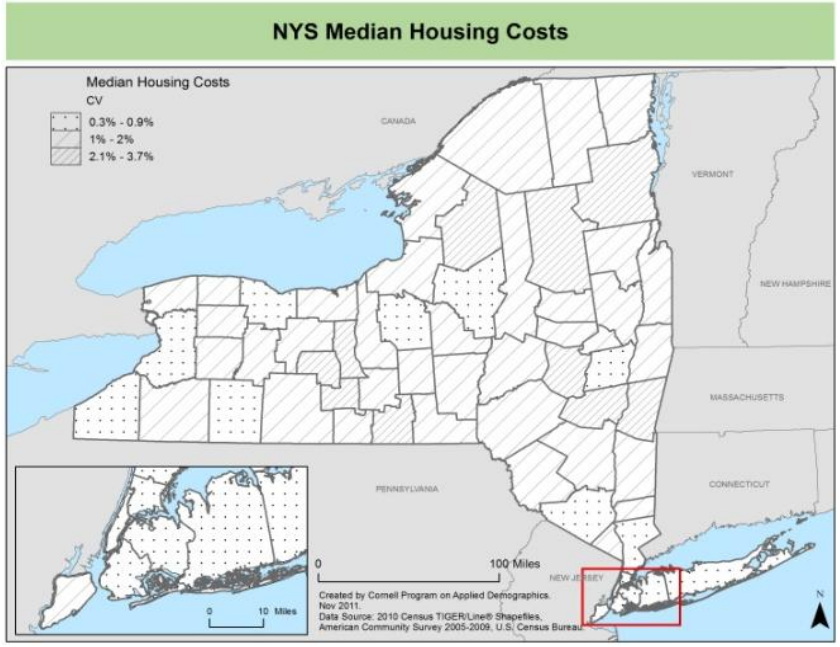


Two Maps or One

- A second major issue is whether to present MOE in separate map or overlay them on the same map and use a “bivariate” legend to aid interpretation.
- Our first comment here is that while experienced users seem to prefer the single, integrated map, casual map readers find both confusing. Need for education
- The second comment pertains to map legends.

Two Maps or One

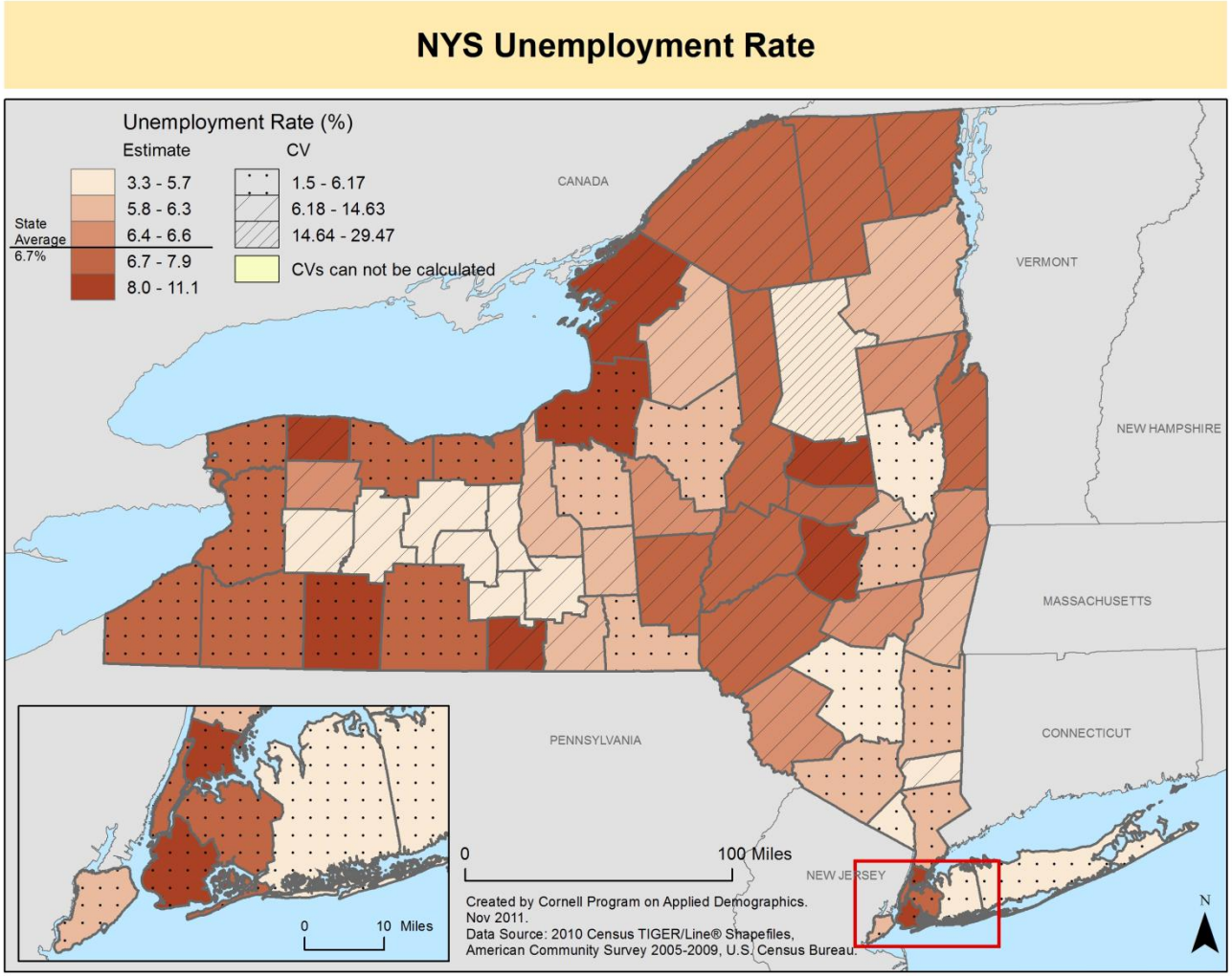




Two Maps or One

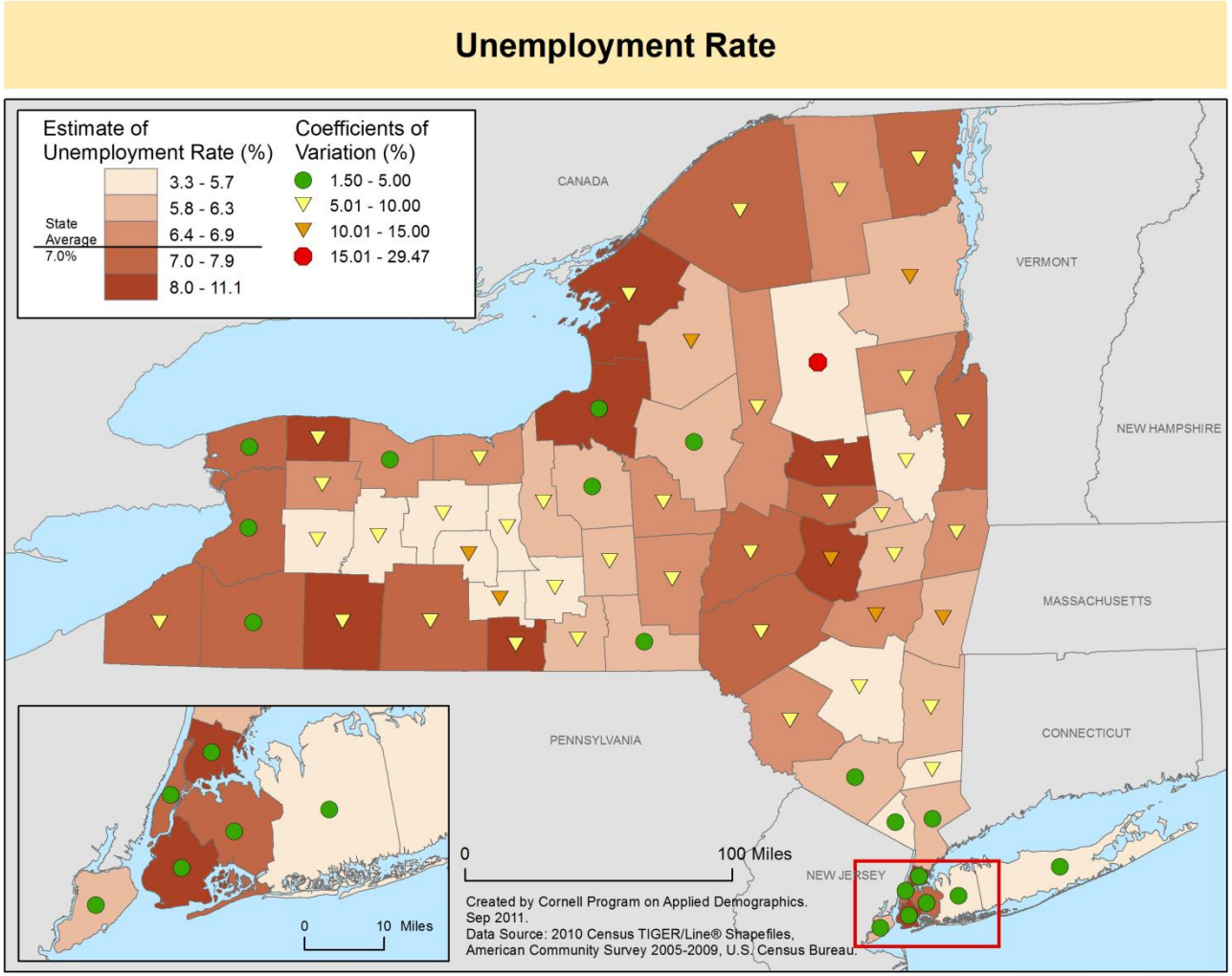
- While not much literature on the topic of legends, we found Wong's ArcGIS extension
  - Too data driven
  - Inflexible methodology (Jenks), break points, # classes
  - Frequently not useful as largest error category was well within bounds of acceptable uncertainty
- We build our own
  - Class breaks at levels decision makers find more useful
  - Flexible methodology and #classes

Symbolizing Uncertainty



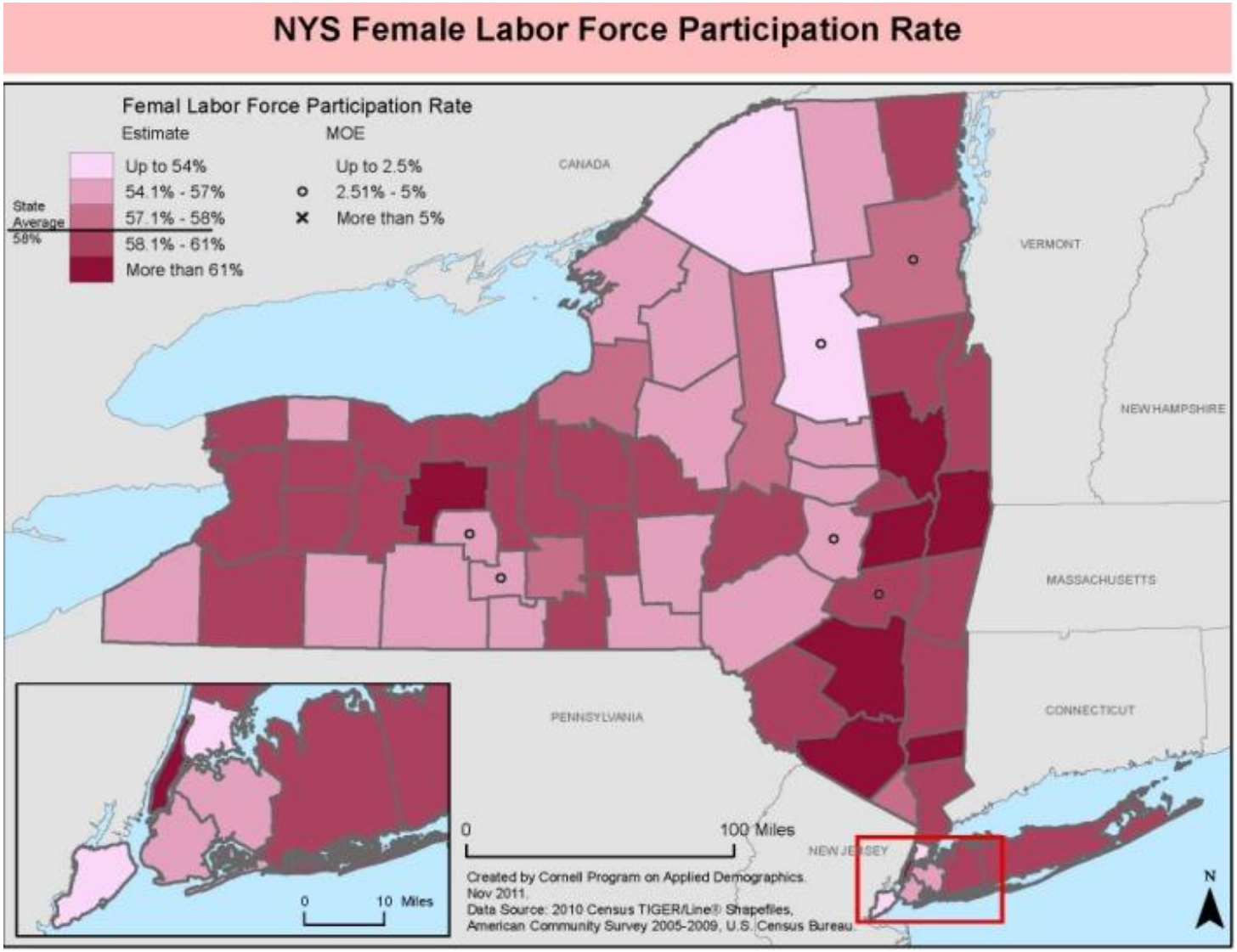
Map 9

Symbolizing Uncertainty



Map 12

Symbolizing Uncertainty



Research on Legends Conducted by Census Bureau

Reliability Legend based on the Coefficient of Variation (CV)

Estimates marked with a blank reliability box (shown below) means that the coefficient of variation (CV) is 0.30 or below, or the estimate is marked as "N" and has been filtered from the table and the CV is not applicable for use.

Estimates marked with a yellow "use caution" reliability box (shown below) means that the coefficient of variation (CV) is above 0.30.

Range	Reliability
CV <=0.30	
CV > 0.30	use caution

The coefficient of variation (CV) is defined as the standard error of an estimate divided by the mean of that estimate, measured as a percentage. Relatively, a lower CV means a more reliable estimate.

Reliability Legend based on the Coefficient of Variation (CV)

Range	Reliability
CV <=0.30	good
0.30 < CV <=0.61	fair
CV > 0.61	poor

The coefficient of variation (CV) is defined as the standard error of an estimate divided by the mean of that estimate, measured as a percentage. Relatively, a lower CV means a more reliable estimate.

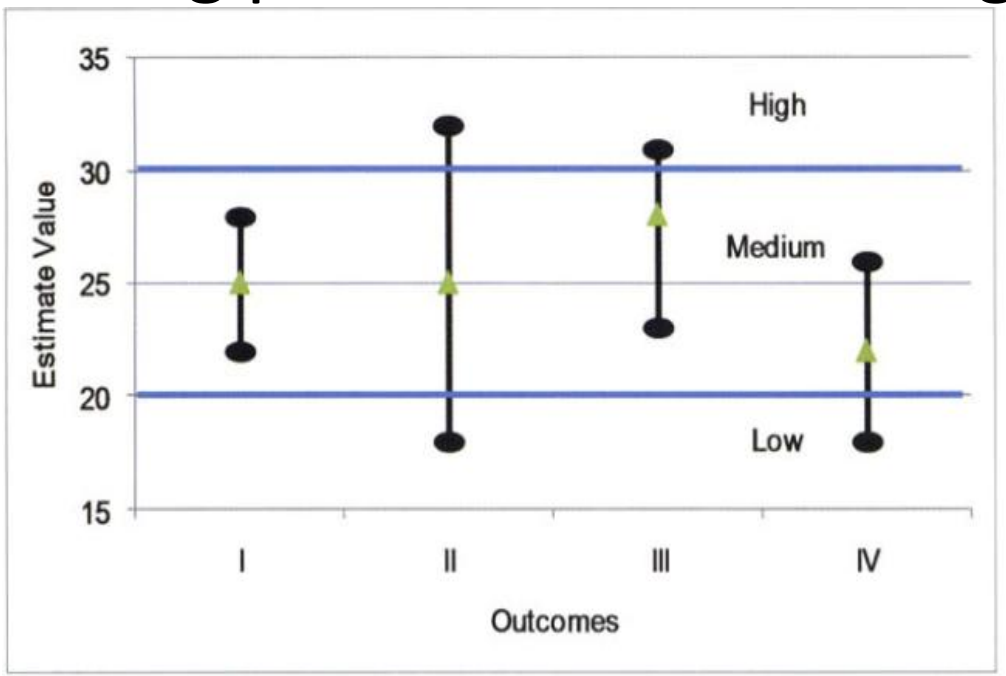
Reliability Legend based on the Coefficient of Variation (CV)

Range	Reliability
CV <=0.10	excellent
0.10 < CV <=0.30	good
0.30 < CV <=0.61	fair
CV > 0.61	poor

The coefficient of variation (CV) is defined as the standard error of an estimate divided by the mean of that estimate, measured as a percentage. Relatively, a lower CV means a more reliable estimate.

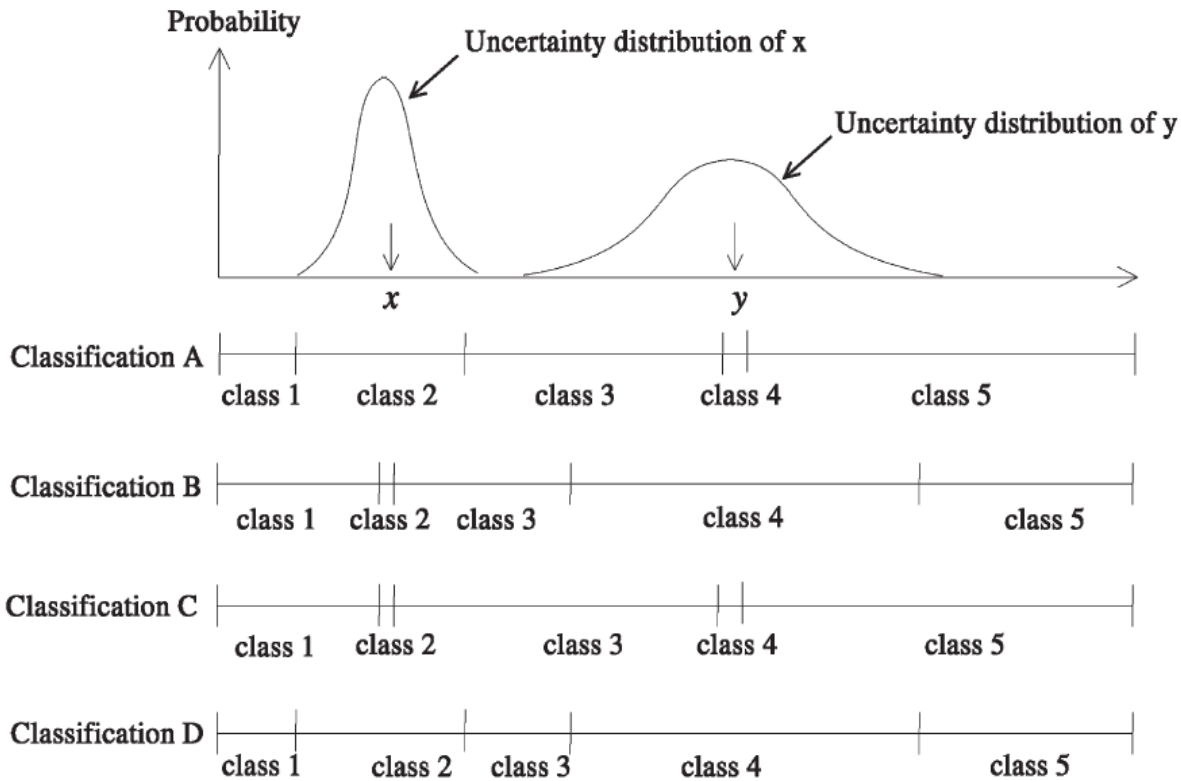
Crisp vs. Fuzzy Classes

- A third issue—employ crisp, sharply defined classes or modified flexible intervals and boundaries in face of uncertainty of estimates.
- Sun and Wong present the issue via graph:



Crisp vs. Fuzzy Classes

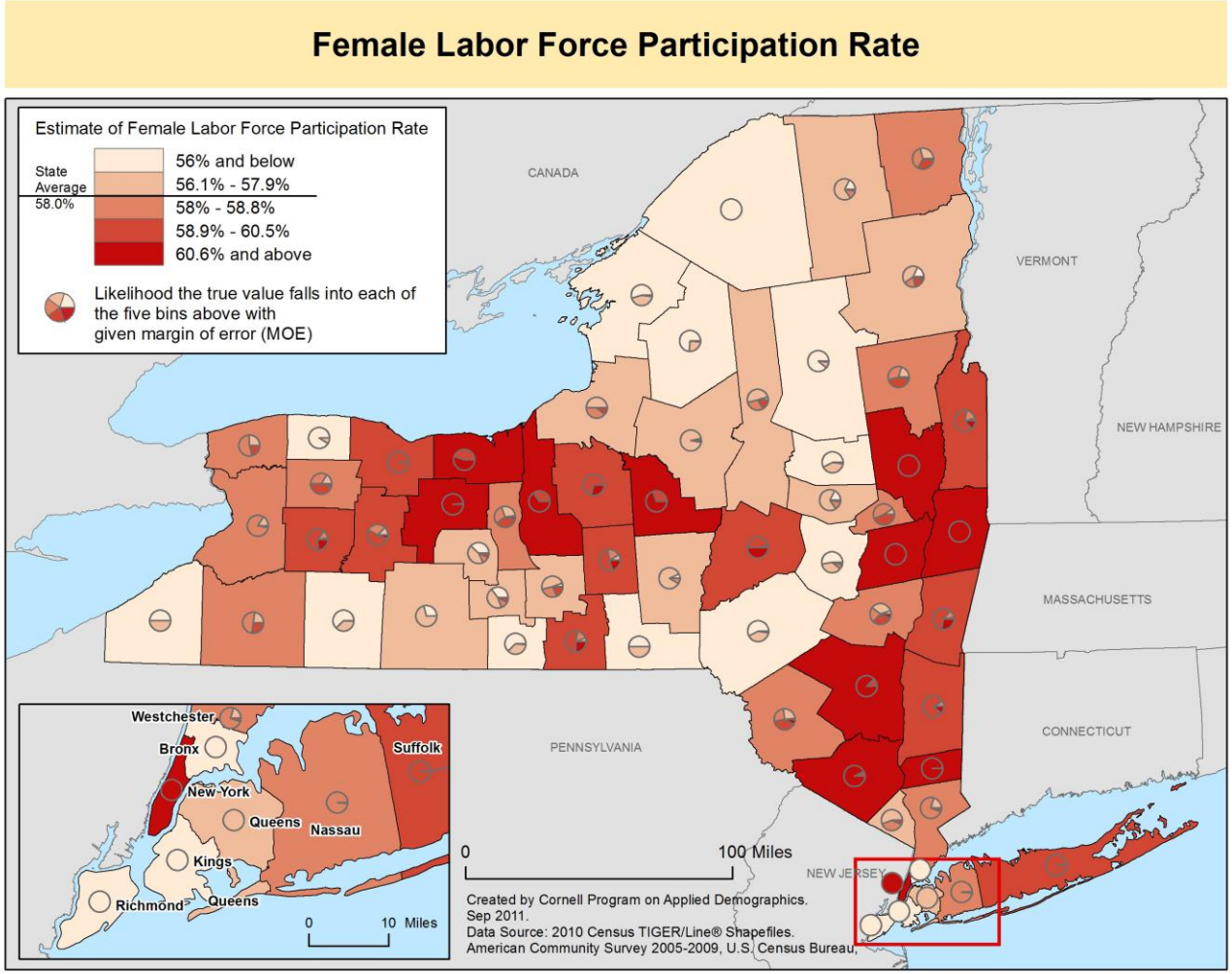
- Xiao et al (2007) present the issue in a slightly different way. They use the term "robustness" to measure how well a classification works



Crisp vs. Fuzzy Classes

- In our own work we explored the idea of portraying the probability that the estimate belonged to the class to which we assign it.
- For static maps we tried the use of pie charts, where each slice of the pie represented the probability that the estimate belonged in the class to which it had been assigned by the Jenks method.

Crisp vs. Fuzzy Classes

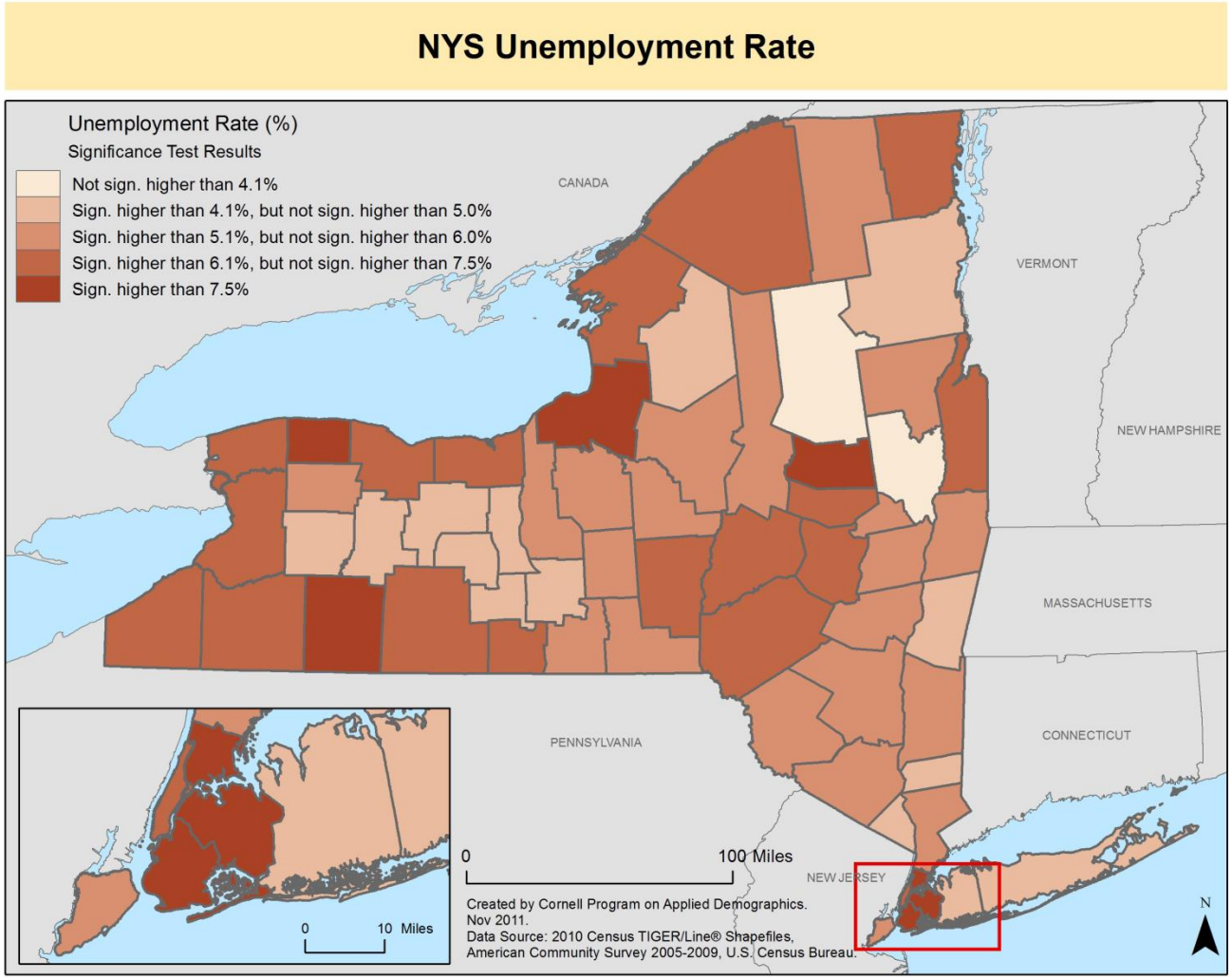


Map 21

Crisp vs. Fuzzy Classes

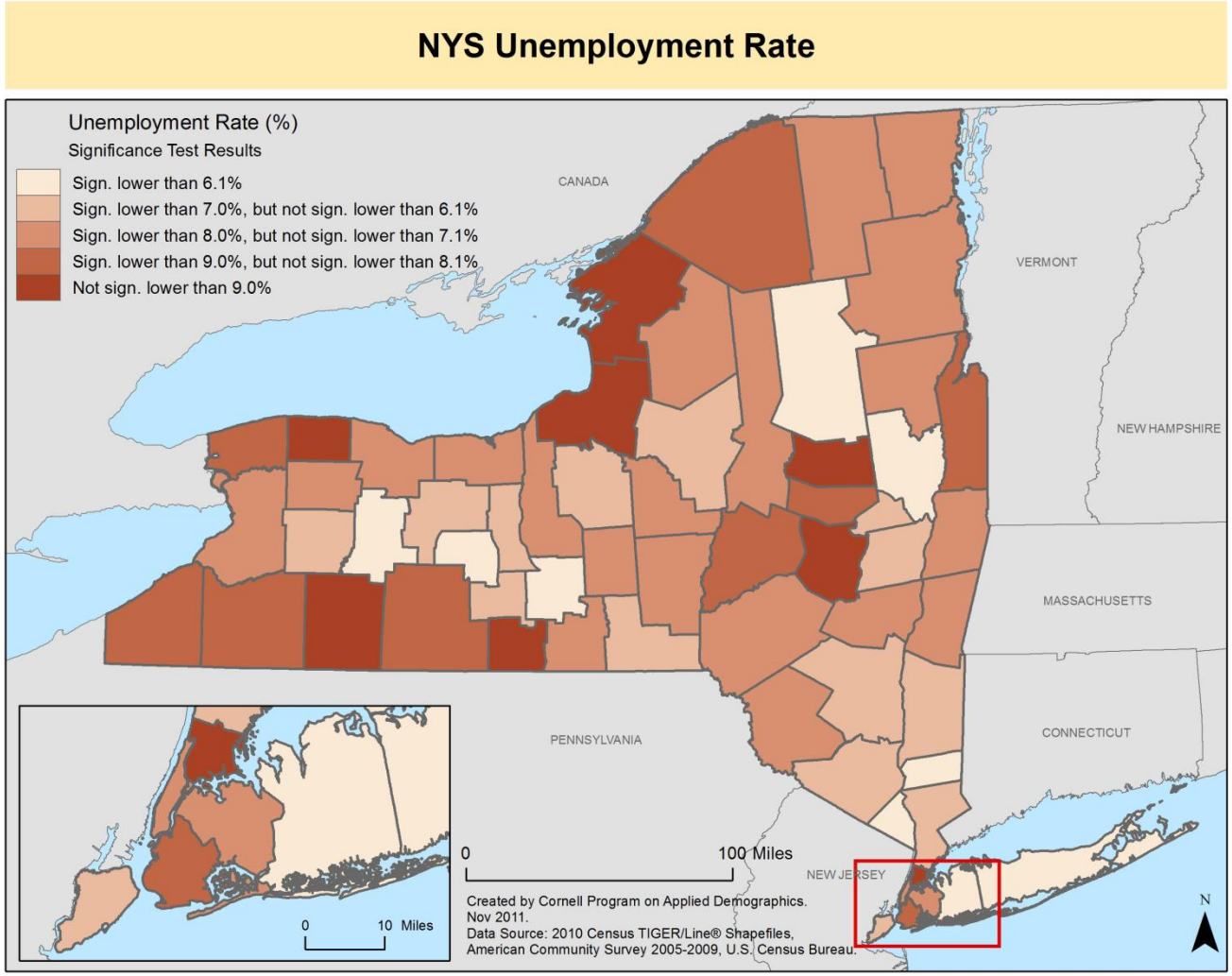
- We also experimented with classifying and displaying the lower bound or the upper bound of the confidence intervals.

Crisp vs. Fuzzy Classes



Map 29

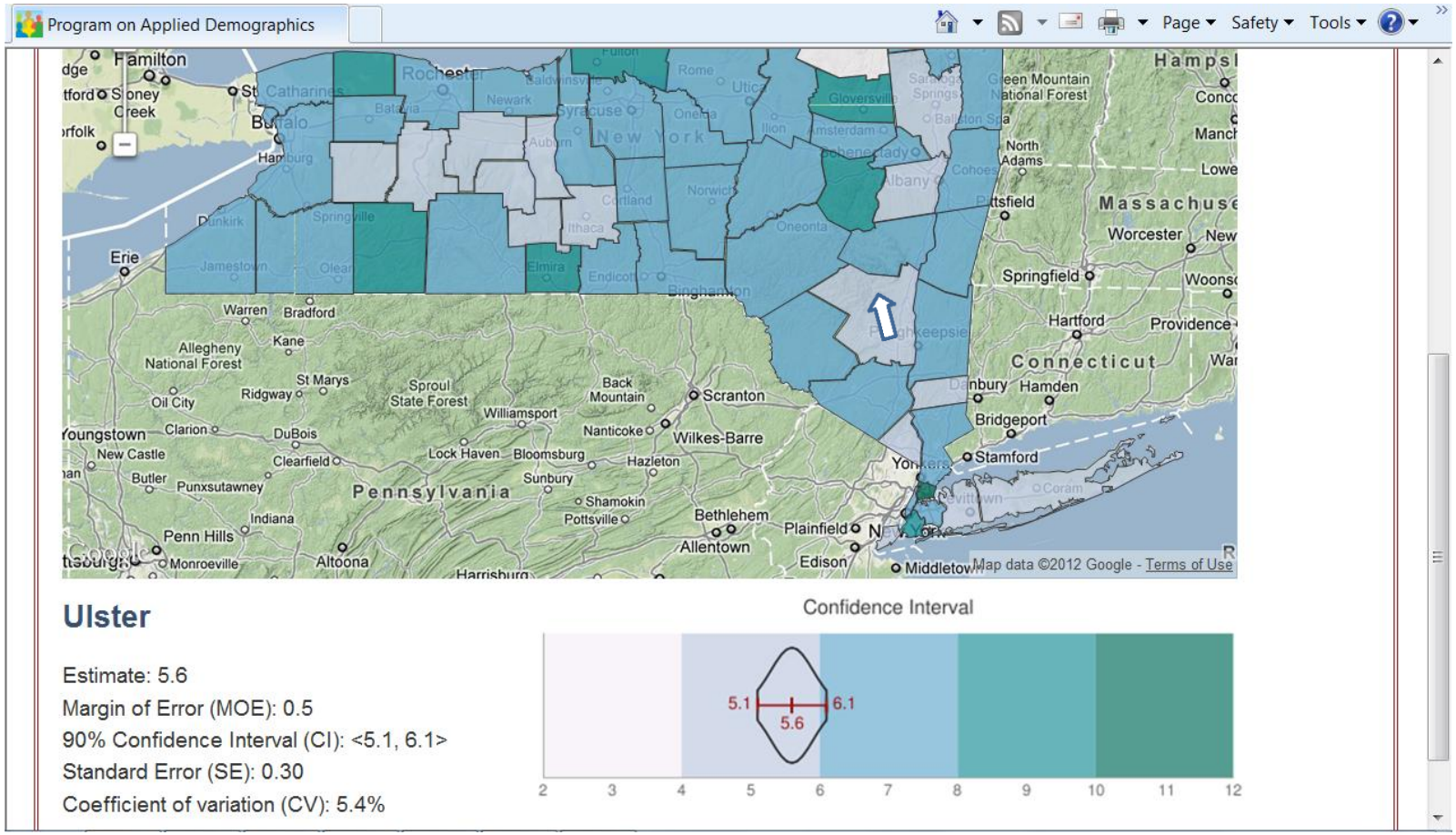
Crisp vs. Fuzzy Classes



Map 30

Crisp vs. Fuzzy Classes

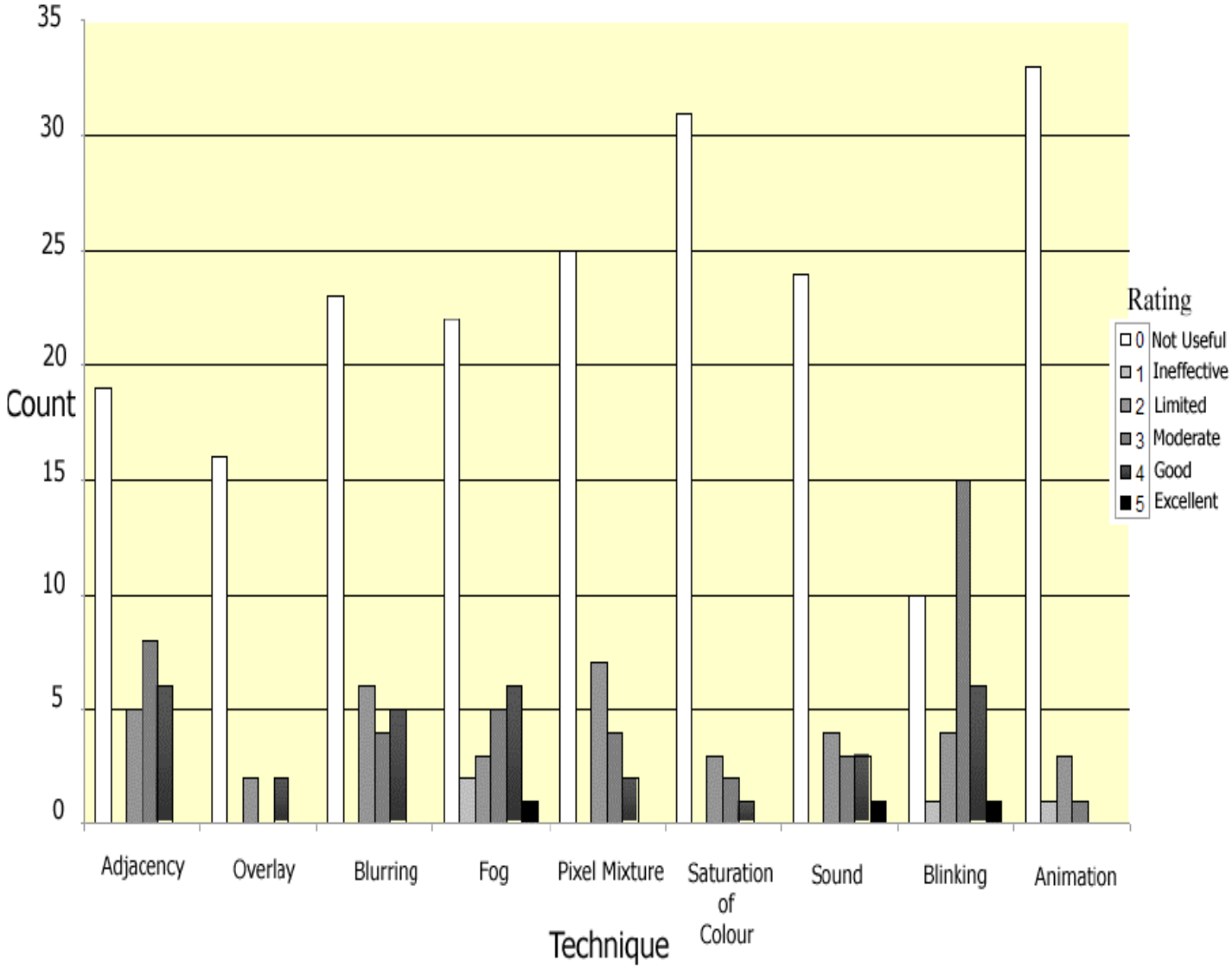
For internet mapping one can provide this information as a feedback when the user clicks a polygon on screen.



Symbolizing Uncertainty

- Kardos, et al researched nine techniques for criteria of usefulness, visual appeal and speed of comprehension.
- They drew the conclusion that the blinking of areas metaphor/technique outperformed the other techniques.
- Overlay was found useful by over 80% of the respondents as was adjacent maps (one for the estimate and one for uncertainty), with “fogging” and “bluring” next most useful.

Symbolizing Uncertainty



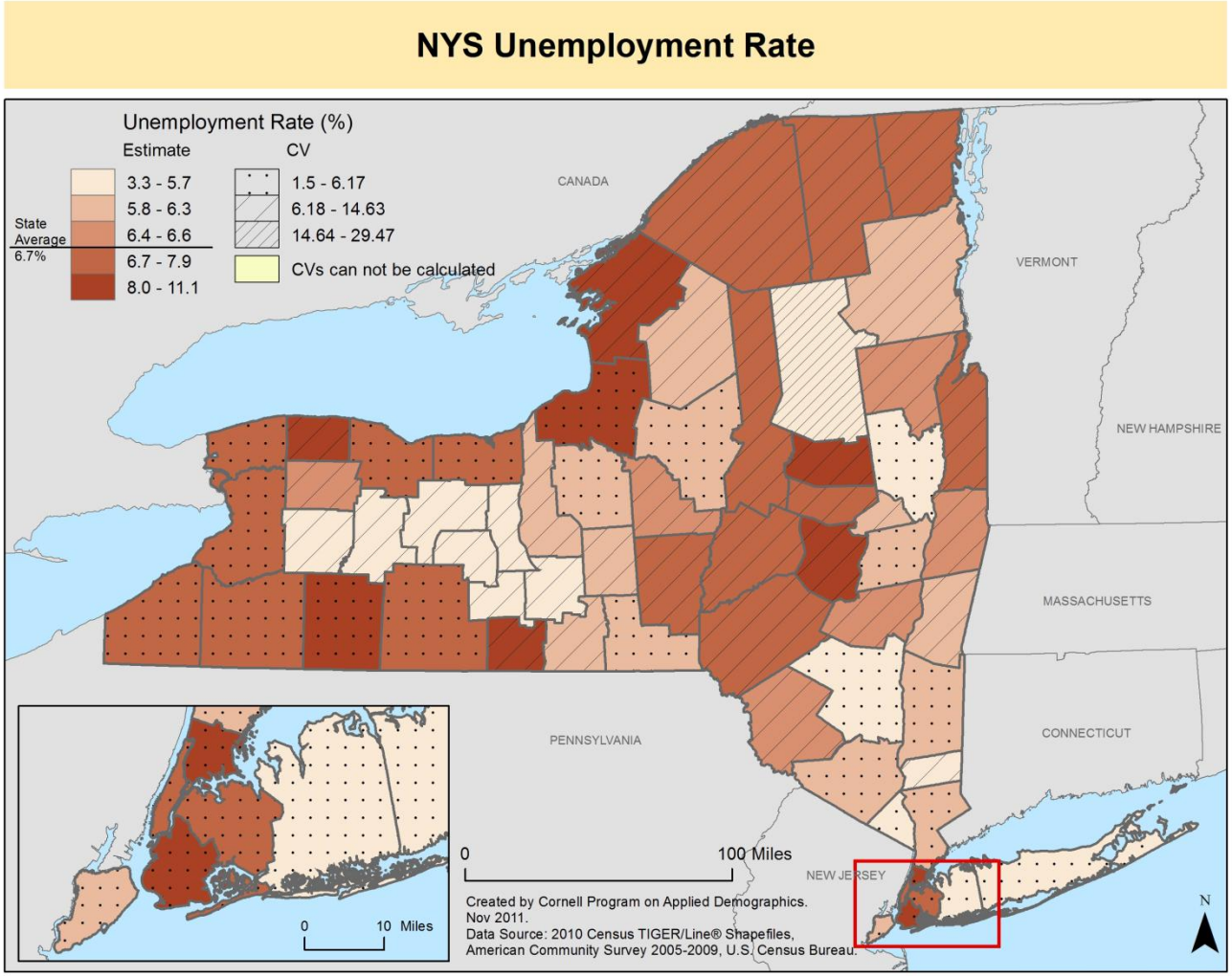
Symbolizing Uncertainty

- The reason for preferring the blinking technique over others was that it didn't obstruct their viewing of the original information values.
- While respondents found the overlay technique useful, the felt it interfered with their understanding of the values symbolized by color.
- We found the same problem of confusion when presenting both the estimate and uncertainty information overlay.

Symbolizing Uncertainty

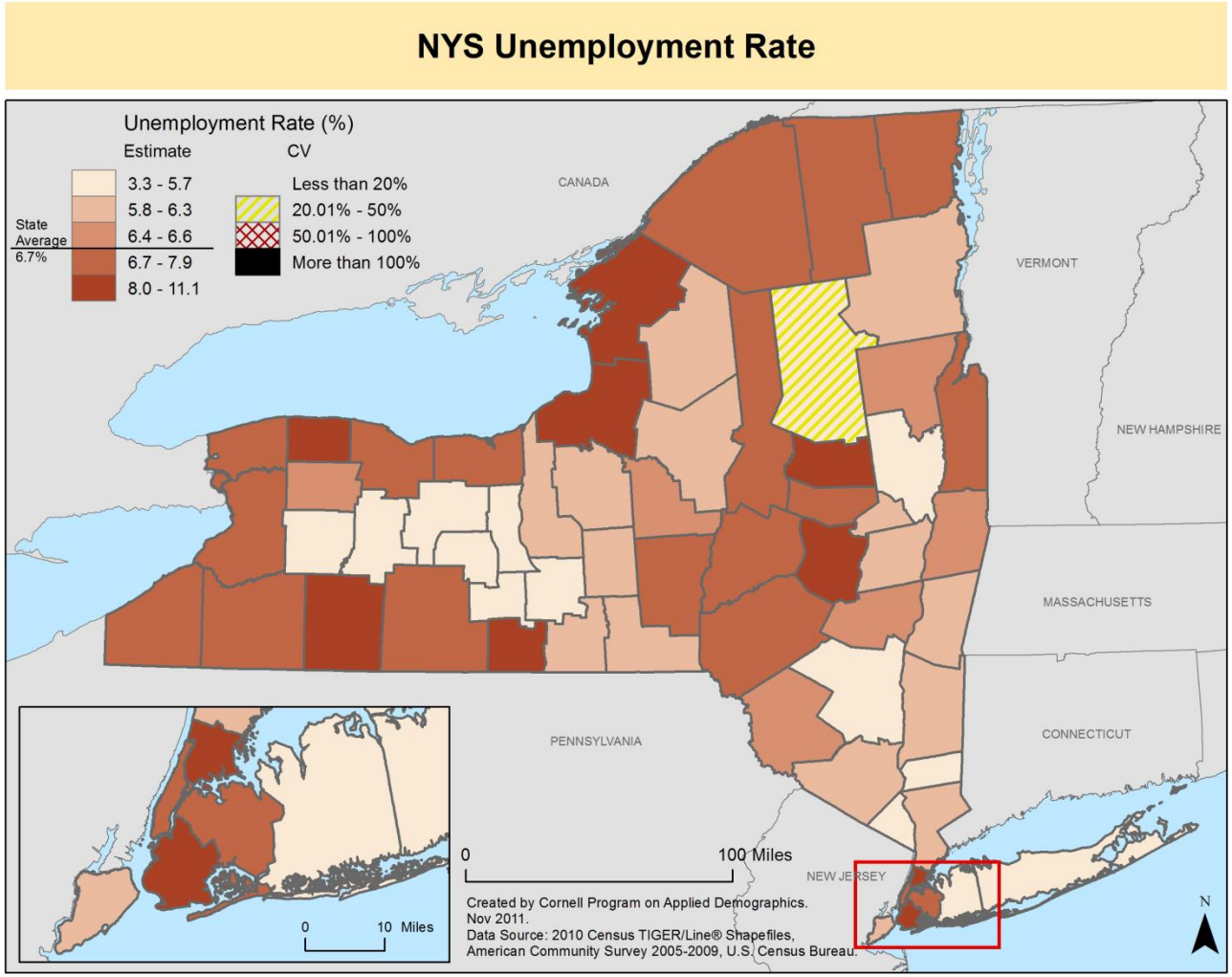
- We experimented with all kinds of symbols—circles, various iconic symbols (filled and unfilled), cross-hatching.
- People seemed to like the cross-hatching the least, stating that it tended to cover up the background estimate information.

Symbolizing Uncertainty



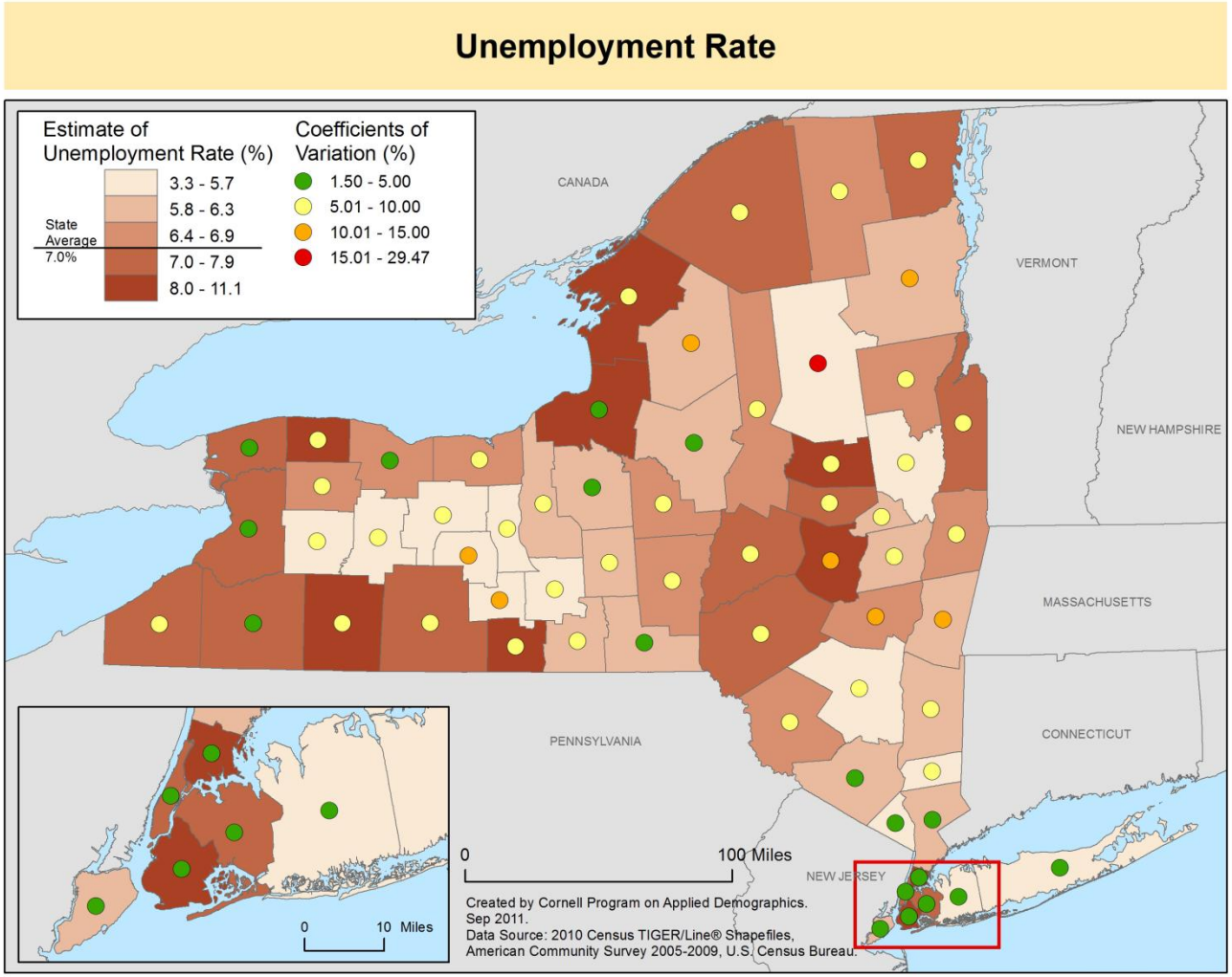
Map 9

Symbolizing Uncertainty



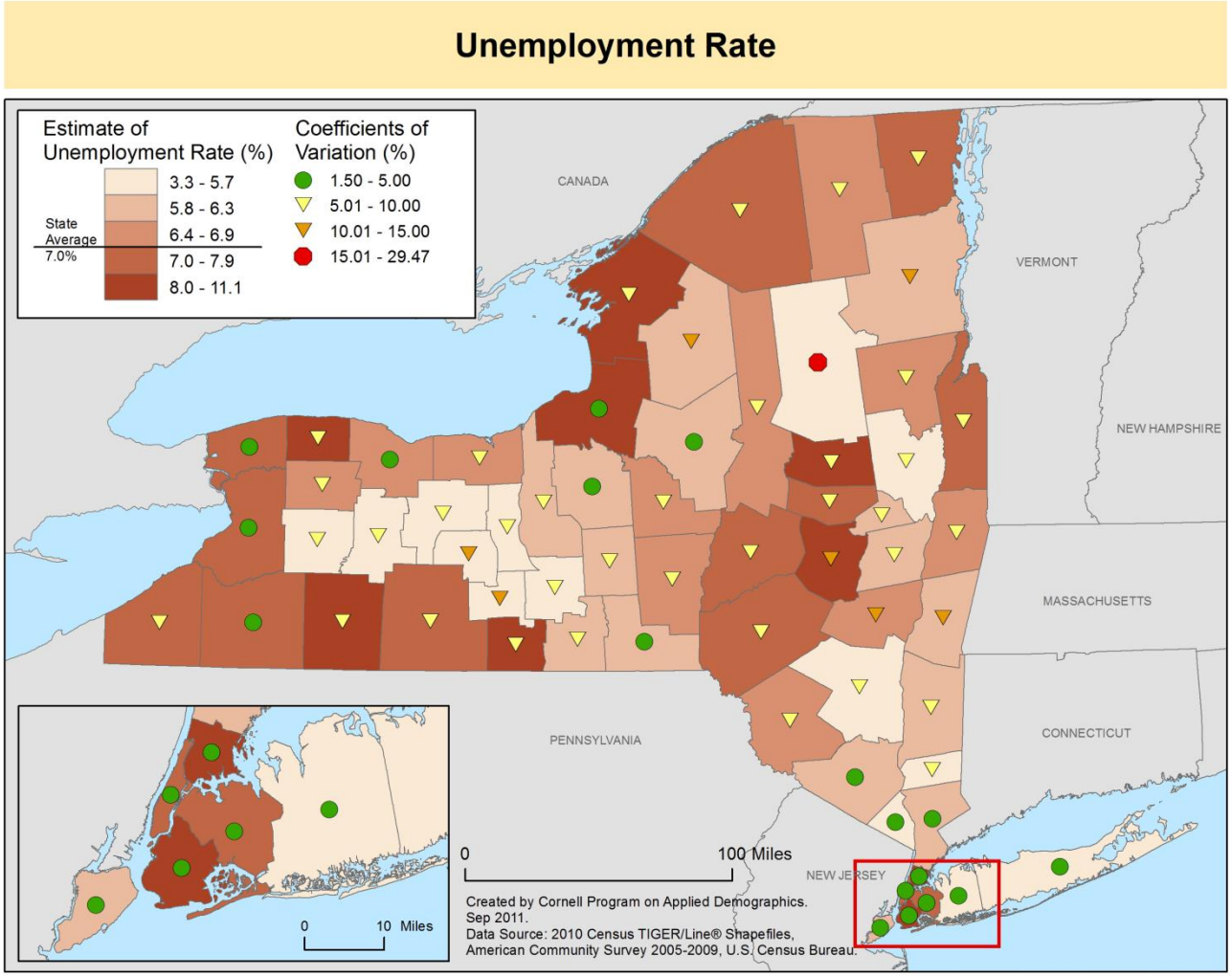
Map 10

Symbolizing Uncertainty



Map 11

Symbolizing Uncertainty

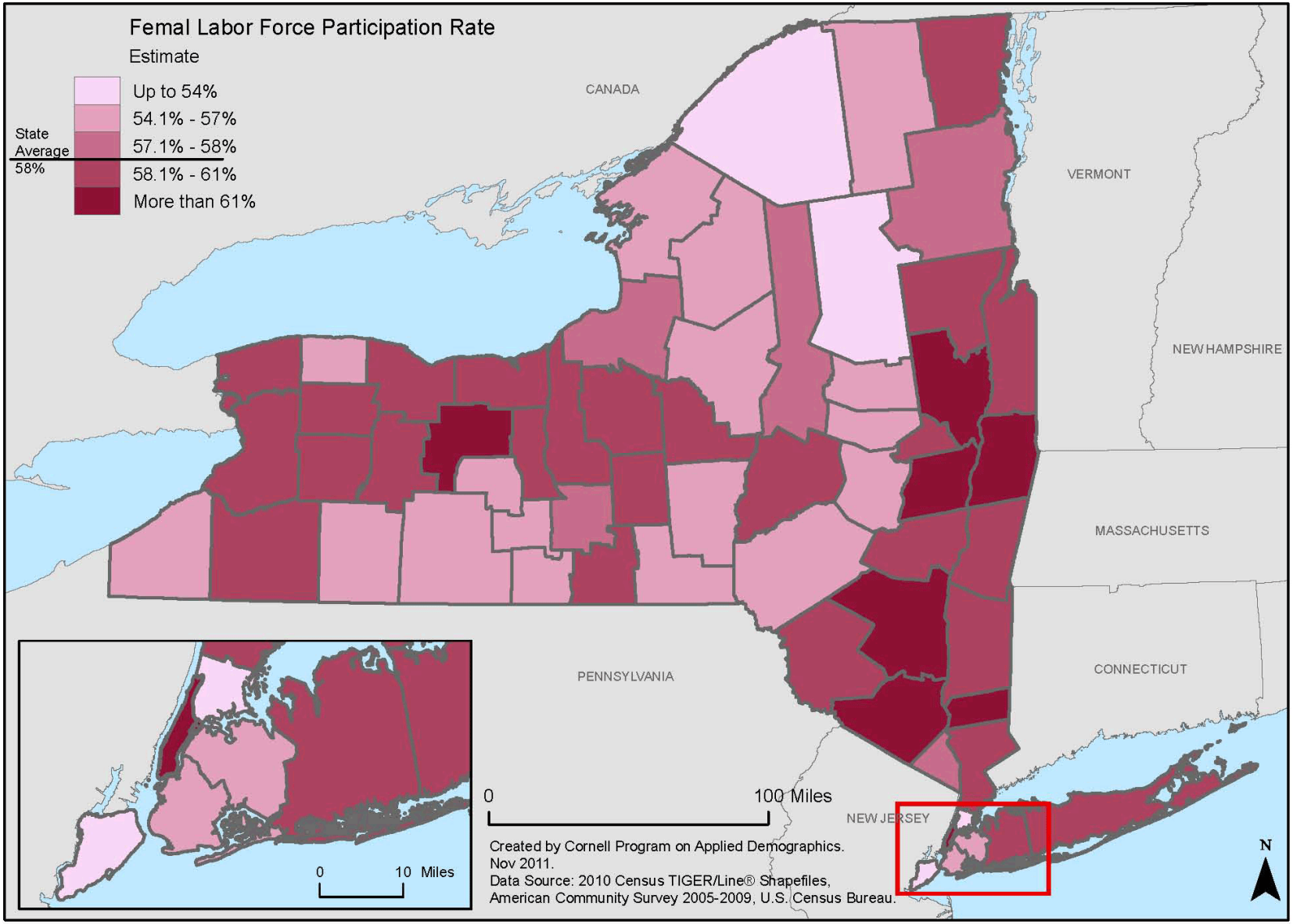


Map 12

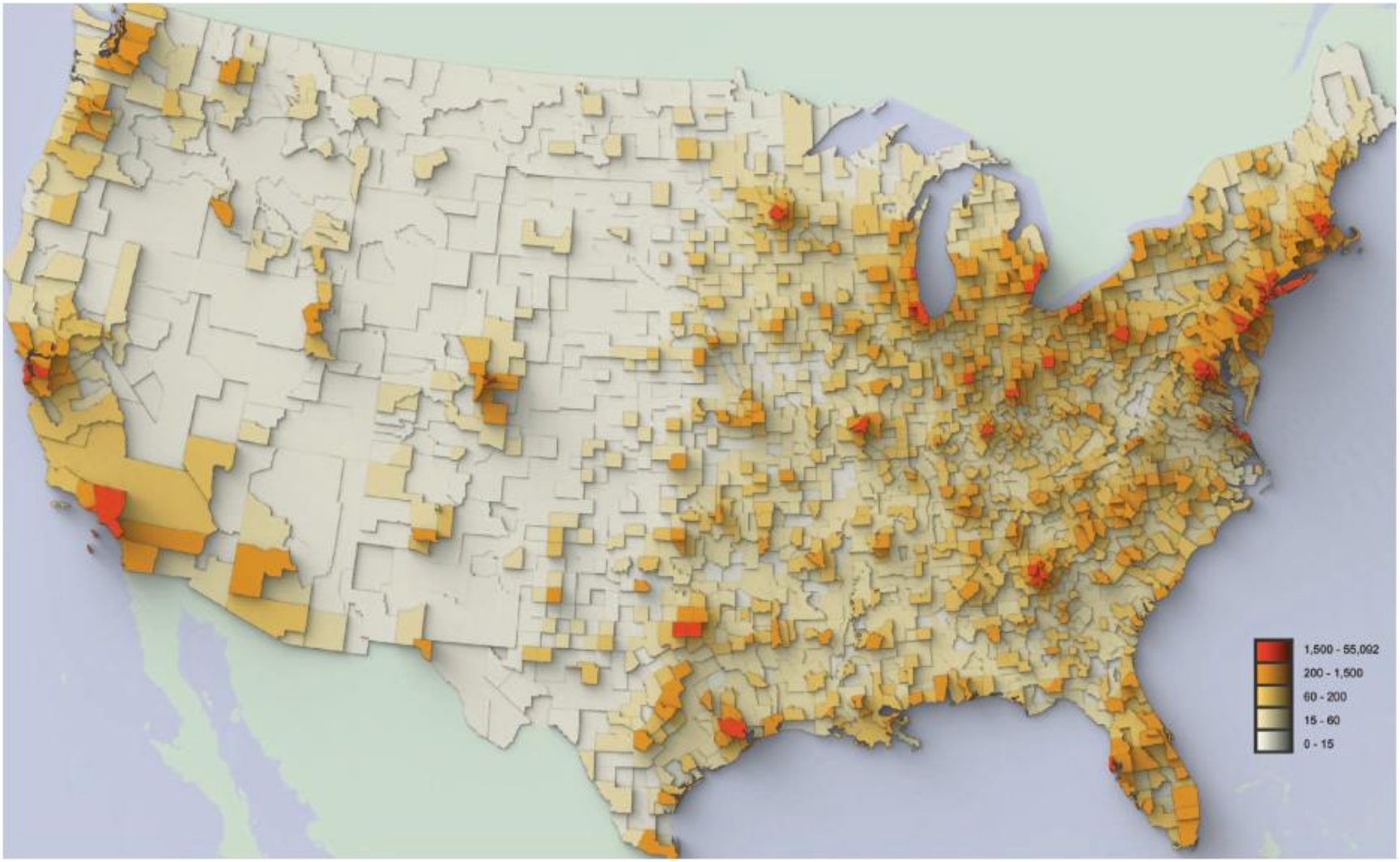
Symbolizing Uncertainty

- We also explored a modification of a “blinking” technique.
- For our static (pdf) maps we first present the estimate for the geographic areas of interest and then, with one mouse click, the viewer overlays the error of estimation information.
- For our interactive, internet maps the user has only to move the mouse over a geographic unit of interest and the error of estimation is displayed.

NYS Female Labor Force Participation Rate



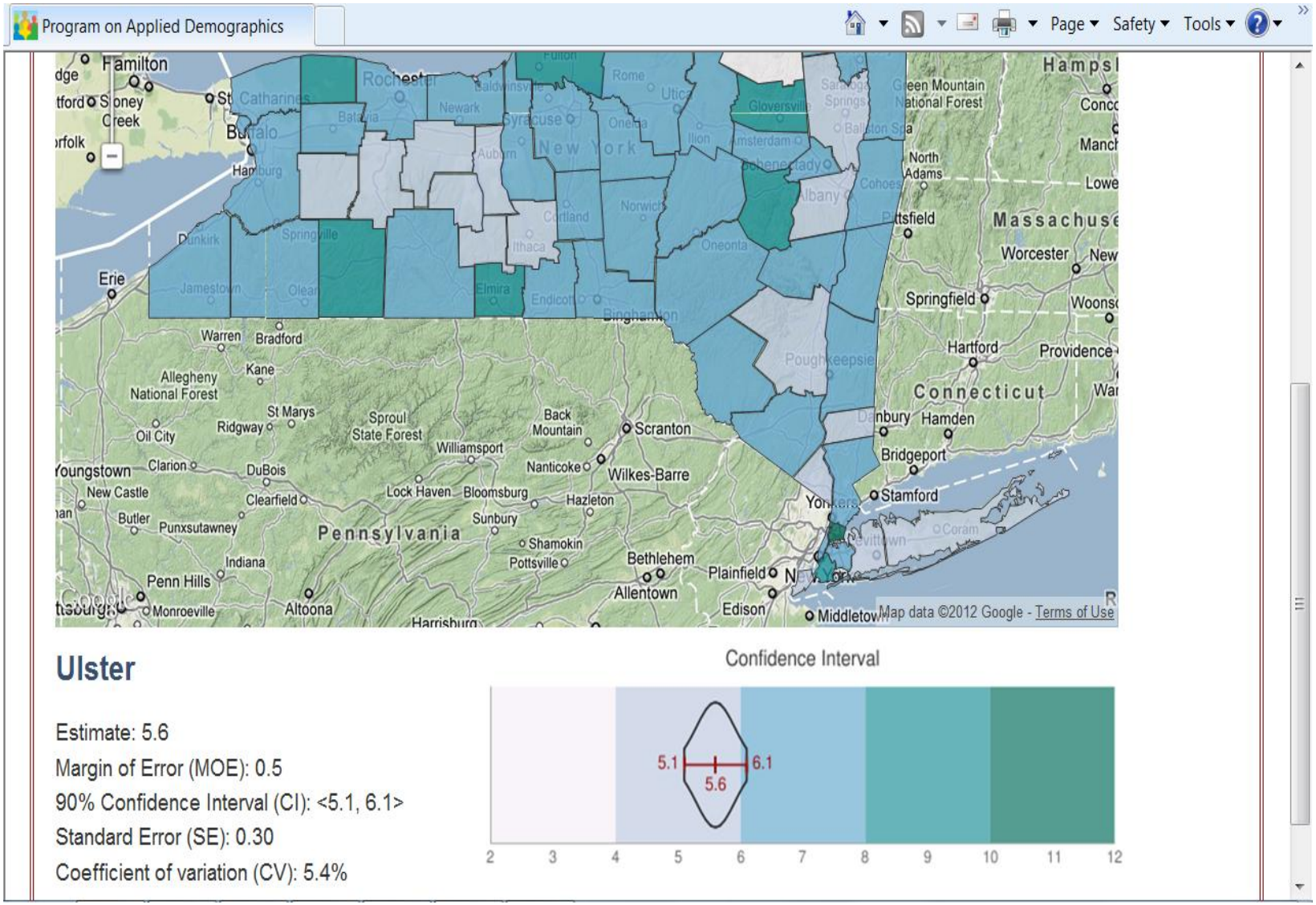
Symbolizing Uncertainty



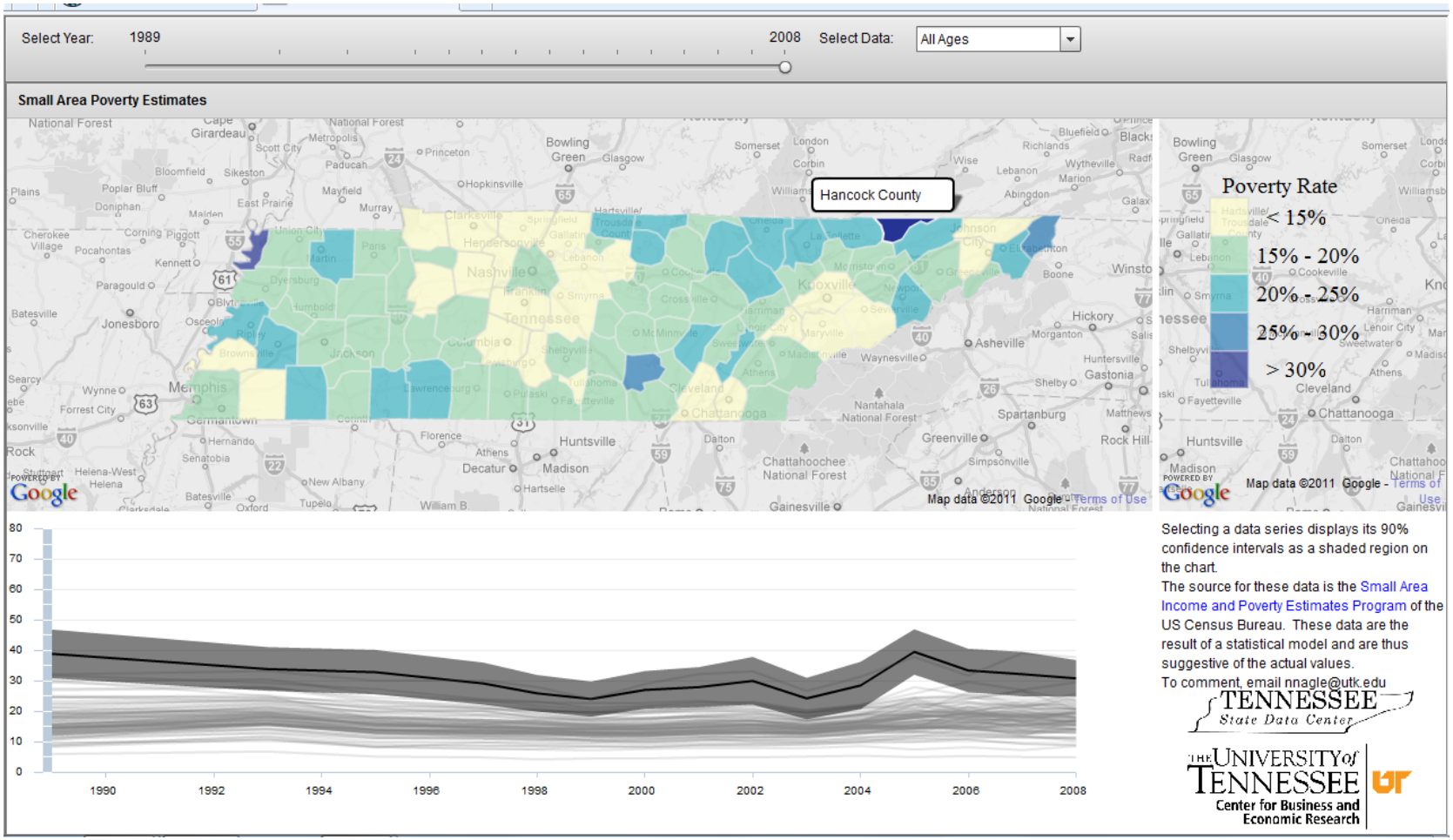
Static vs. Online Maps

- Dynamic interactive maps permit much more flexibility in presenting information compared to static maps. With a slider that lets you choose the value to compare to:
- The one I showed earlier, giving detailed information on uncertainty in each county:
- http://pad.human.cornell.edu/papers/annex/uncertaintymap_fullinfo.cfm
- Nagel's internet map for SAIPE data:

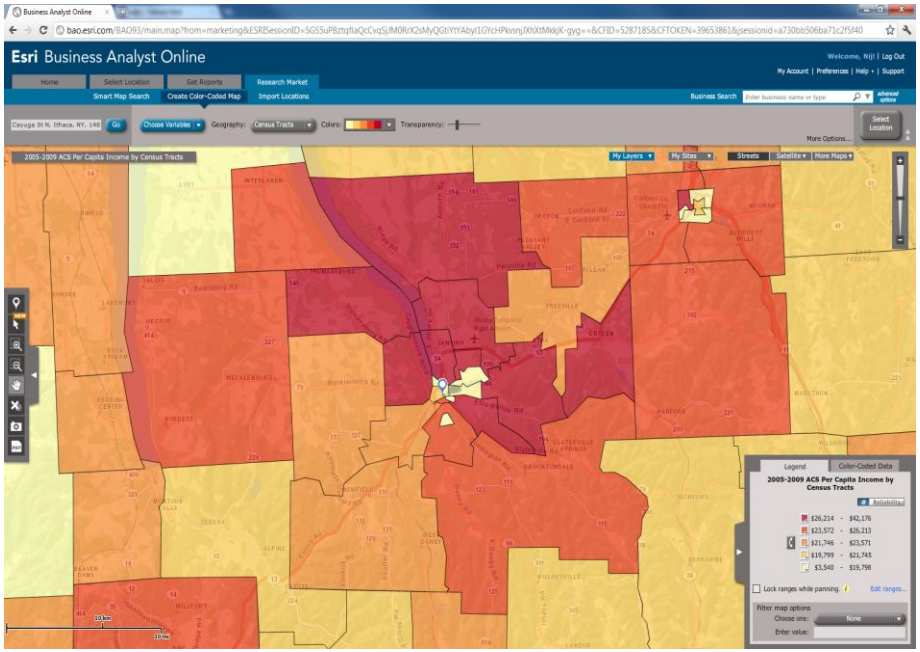
Static vs. Online Maps

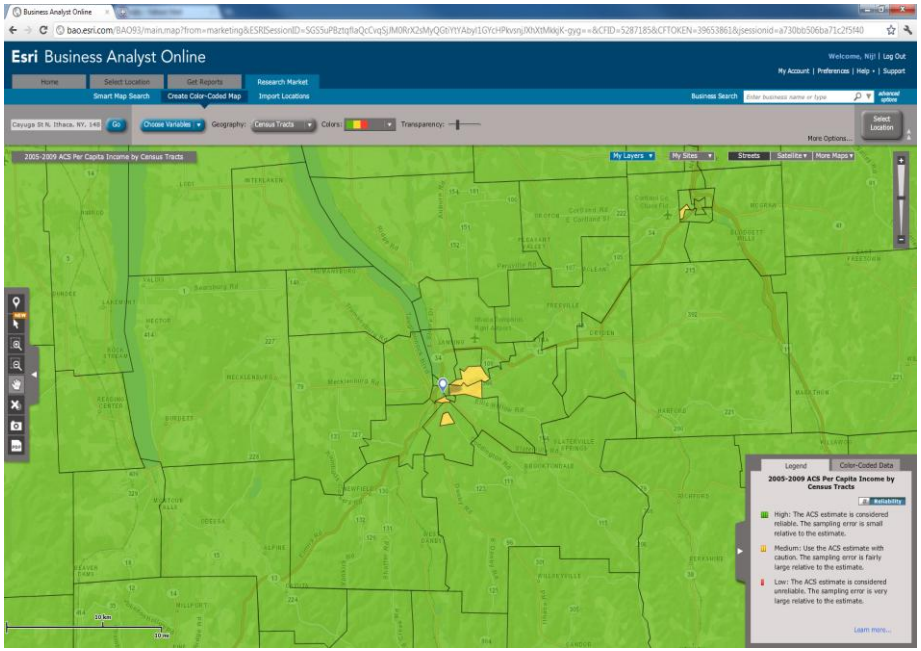


Static vs. Online Maps



Static vs. Online Maps

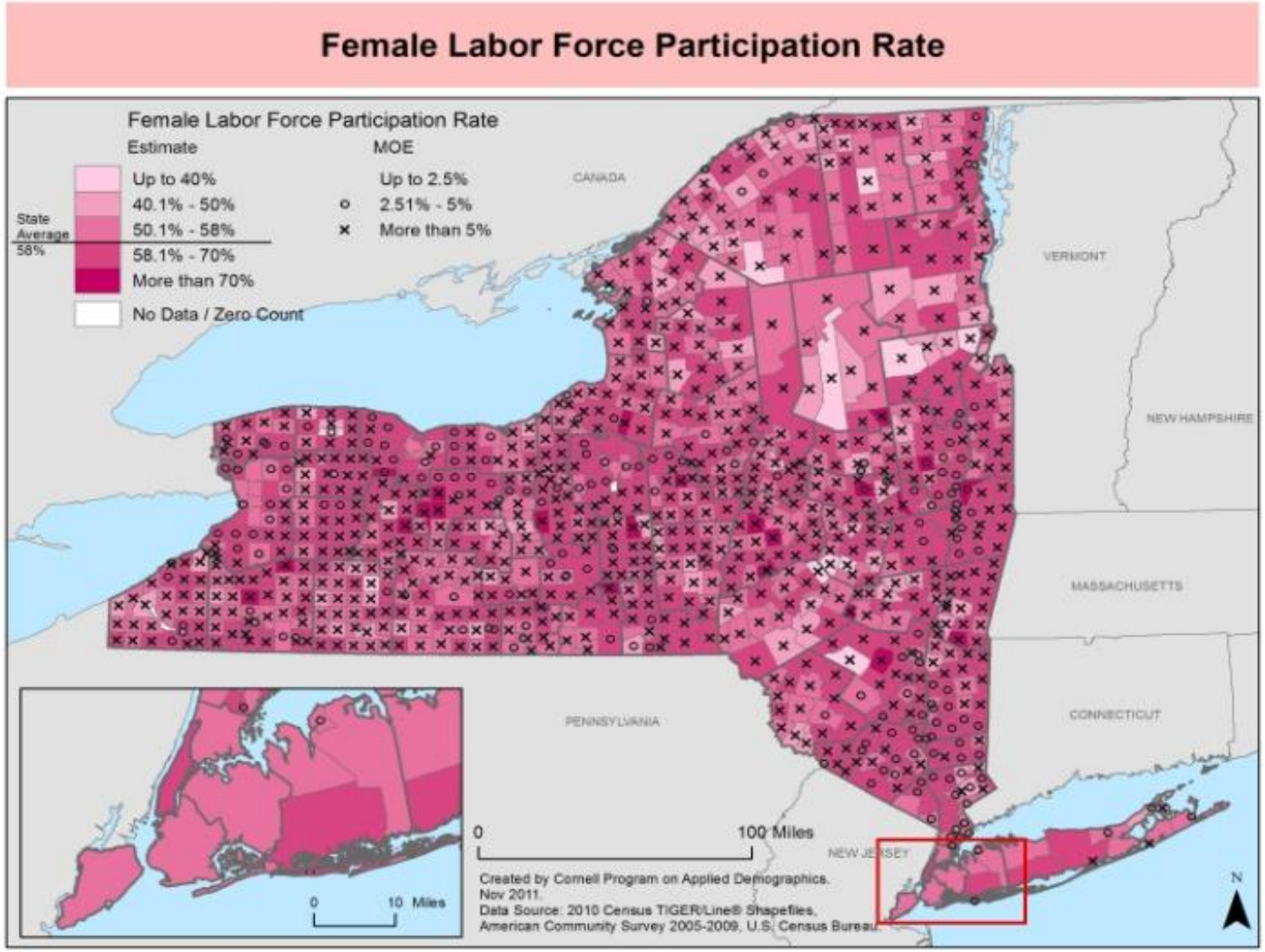




Number of Geographic Units

- Another of the major issues is the number of geographic units presented on the maps and confusion that causes in viewing and interpretation.
- Torrieri et al raise this issue.
- In our own work we find the same problem.
- Compare the next two maps, one at the county level and another at the sub-county level.

Number of Geographic Units



Number of Geographic Units