Disclosure Avoidance and Analytical Validity in "On The Map"

Fredrik Andersson

February 1, 2007

U S C E N S U S B U R E A U

Disclosure Avoidance

Bayesian statistical techniques to create a *partially synthetic* version of the confidential data

- Block of origin counts sampled from a *posterior predictive distribution* conditional on destination block and worker characteristics (earnings, industry, age, ownership sector)
- The *posterior predictive distribution* is derived from combining the *likelihood* ("true data") with a *prior*

So, what does this really mean???

Key Implication

The relative weight of the prior when sampling from the posterior distribution is inversely related to the size of the population being synthesized

- For larger populations the synthetic place of residence data closely mimic underlying data
- For small populations the synthetic place of residence data are relatively more "noisy" to protect confidentiality

Important to keep in mind when making inferences using OnTheMap

How "noisy" an estimate is can be assessed by taking advantage of all 10 implicates of the synthetic data available on the virtual RDC

Level of protection increases as population in work block decreases

Mean proportion of workers that need to be reallocated across selected residence areas in the synthetic data to replicate confidential data

Population in Work Block	Counties	Census Tracts	Blocks
1-5	30%	36%	43%
6-10	23%	25%	29%
11-20	18%	23%	24%
21-50	12%	18%	19%
51-100	10%	15%	17%
101-250	6%	11%	13%
250-500	5%	9%	13%
501-high	3%	7%	11%