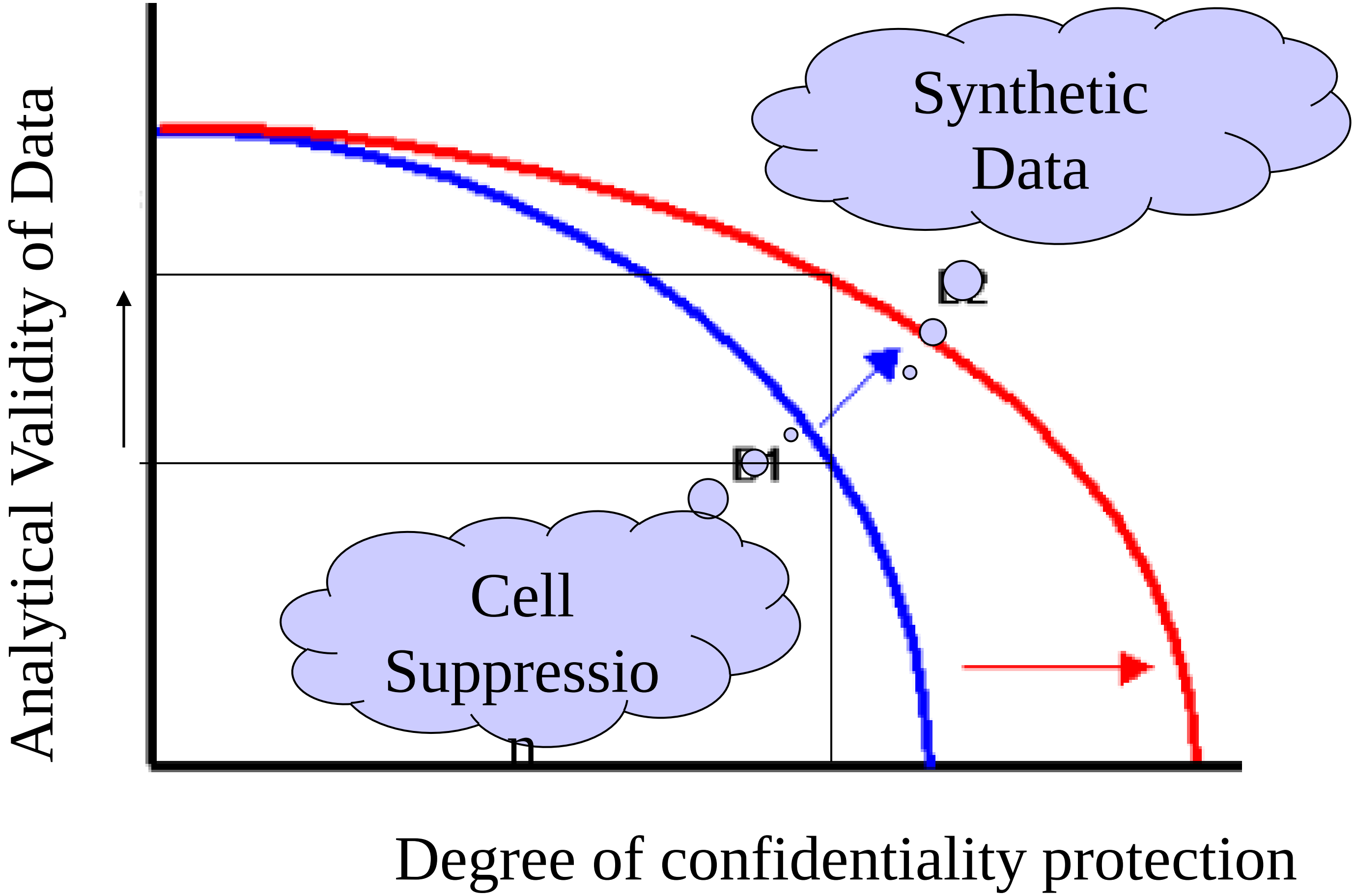
Disclosure Avoidance and Analytical Validity in “On The Map”

Fredrik Andersson

February 1, 2007

The Challenge: *Maximize Analytical Validity of Data Subject to Strict Confidentiality Protection Constraints*



Goals of Presentation

Present the Disclosure Avoidance Protocol for *On The Map*

Demonstrate that synthetic data methods can be used to protect confidentiality while preserving analytical validity

Basic Facts about the Disclosure Protection System for OnTheMap

Goal: "to protect confidentiality while preserving analytical validity of data"

- No cell suppression
- Synthetic place of residence data
- Workplace data protected by QWI disclosure protection system ("dynamically consistent noise infusion")

First-ever data product released by a Statistical Agency (Feb 2006) that relies on synthetic data method as its primary disclosure avoidance technique

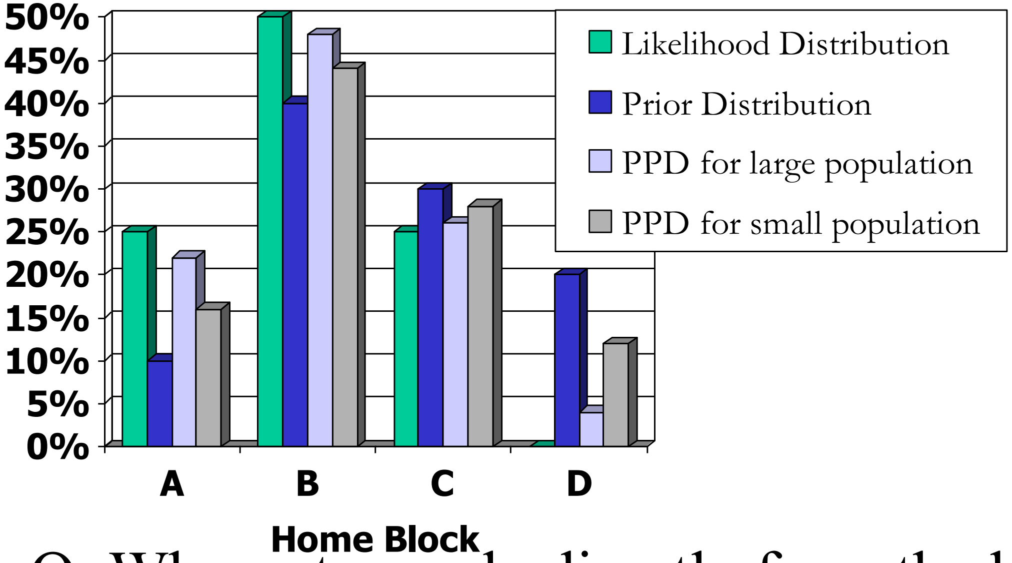
Disclosure Avoidance

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???

Creation of Synthetic Data



Fictional example: Distribution of place of residence for workers in a specific block, industry, earnings category, age category, ownership secto

Q: Why not sample directly from the likelihood/What’s the role of a prior?

Q: How are the priors constructed?

Q: How much weight is given to the prior?

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

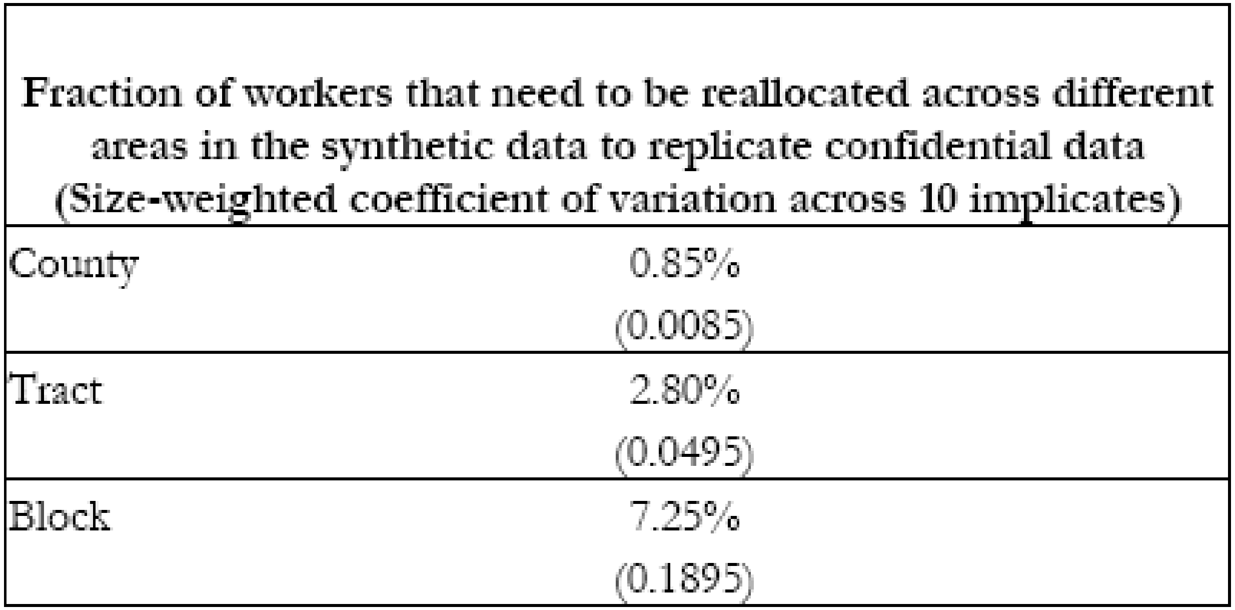
- For larger populations the synthetic place of residence data closely mimic underlying data
- For small populations the synthetic place of residence data are relatively more "noisy" to protect confidentiality

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

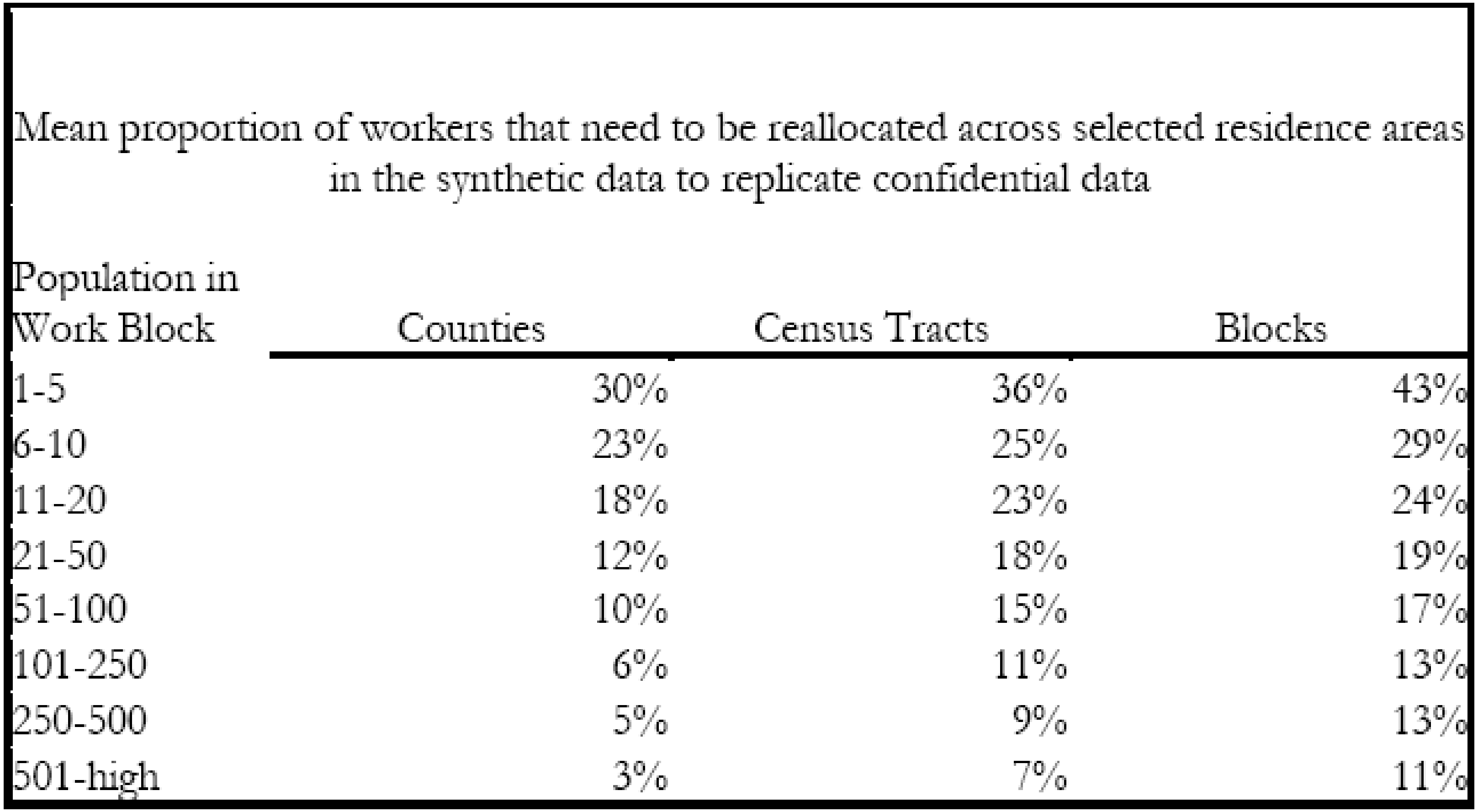
Empirical Results on Analytical Validity & Confidentiality Protection

The residence patterns in synthetic data mimic confidential data well



	Fraction of workers that need to be reallocated across different areas in the synthetic data to replicate confidential data (Size-weighted coefficient of variation across 10 implicates)
County	0.85% (0.0085)
Tract	2.80% (0.0495)
Block	7.25% (0.1895)

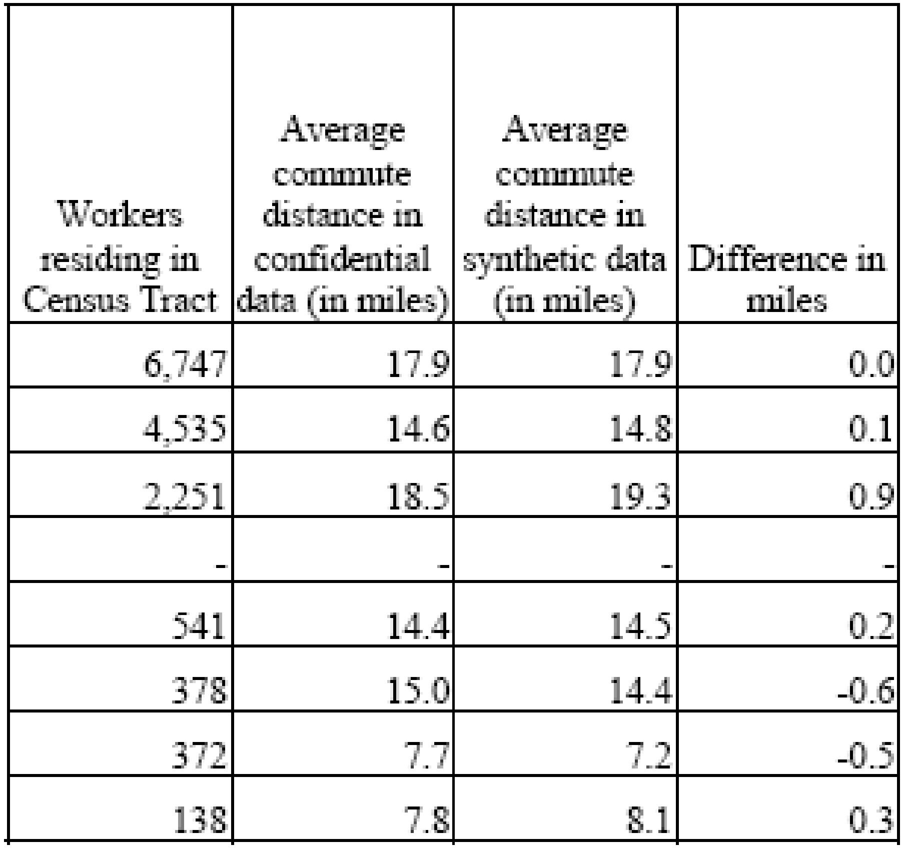
Level of protection increases as population in work block decreases



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%

Key Properties in data, such as commute distance, are preserved in synthetic data



Workers residing in Census Tract	Average commute distance in confidential data (in miles)	Average commute distance in synthetic data (in miles)	Difference in miles
6,747	17.9	17.9	0.0
4,535	14.6	14.8	0.1
2,251	18.5	19.3	0.9
-	-	-	-
541	14.4	14.5	0.2
378	15.0	14.4	-0.6
372	7.7	7.2	-0.5
138	7.8	8.1	0.3