Synthetic Data for Small Area Estimation and Other Geographical Analyses

Trivellore Raghunathan (Raghu)

Department of Biostatistics and Institute for Social Research (ISR), University of Michigan

(Collaborator: Mandi Yu, Graduate Student in the Michigan Program for Survey Methodology, ISR)

August 21, 2007

IAB-Nuremberg, Germany

Motivation

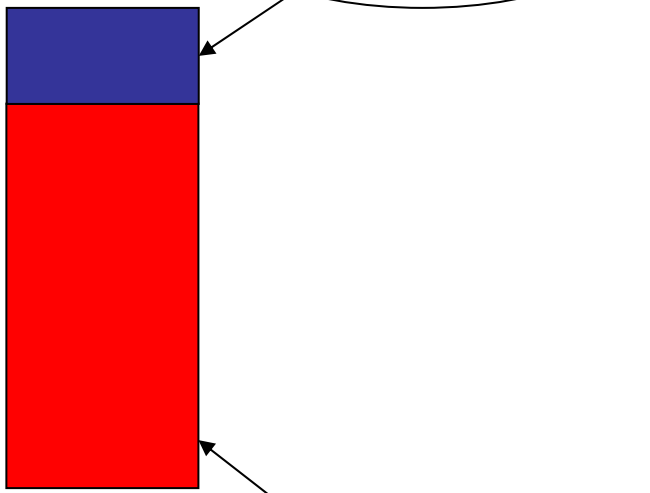
- Increasing demands for estimates for subnational estimates such as states and provinces, counties, school districts etc.
- Increase in demands for geographical details to permit epidemiological analyses of neighborhood level factors affecting individual health
- Releasing of such geographical details may increase the risk disclosure

Goals

- Procedures for generating synthetic data sets with enough geographical details to permit needed analysis
- Evaluate inferences from synthetic data sets by comparing them to those based on actual data sets
  - This talk considers simulated data sets
  - Work is in progress using the actual data sets
    - A possible candidate is the American Community Survey

Framework for analyzing survey data

- Approach exploits the Bayesian and missing data framework for inference from surveys



Sampled subjects

Non-sampled subjects

Bayesian inference:

$\Pr(y_{ns} \mid y_s)$

$\Pr(Q(y_s, y_{ns}) \mid y_s)$

Monte – Carlo implementation:

Draw $y_{ns}^{*(l)}$ *from* $\Pr(y_{ns} \mid y_s), l = 1, 2, \ldots,$

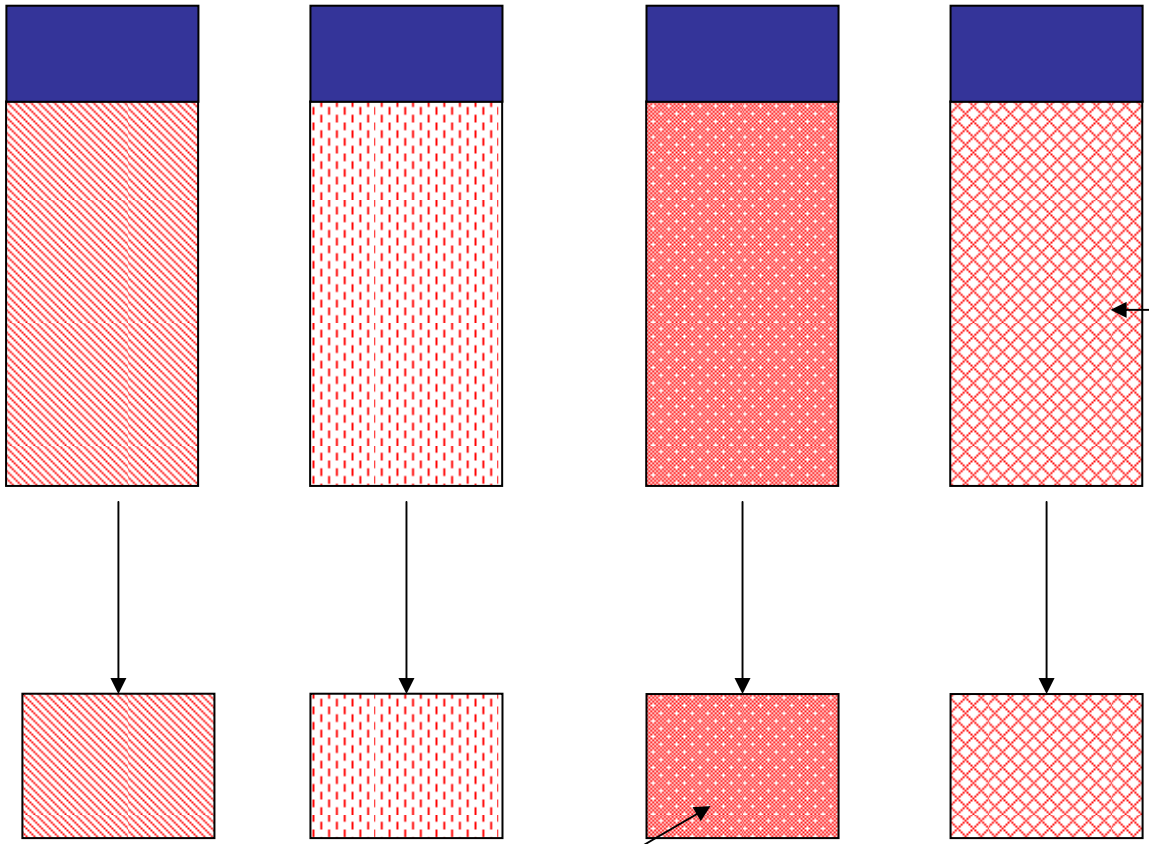
$Q(y_s, y_{ns}^{*(l)}), l = 1, 2, \ldots.$ *are draws from*

$\Pr(Q(y_s, y_{ns}) \mid y_s)$

Multiple imputation:

Small number of draws

Multiple Synthetic Samples (Rubin, 1993, JOS)

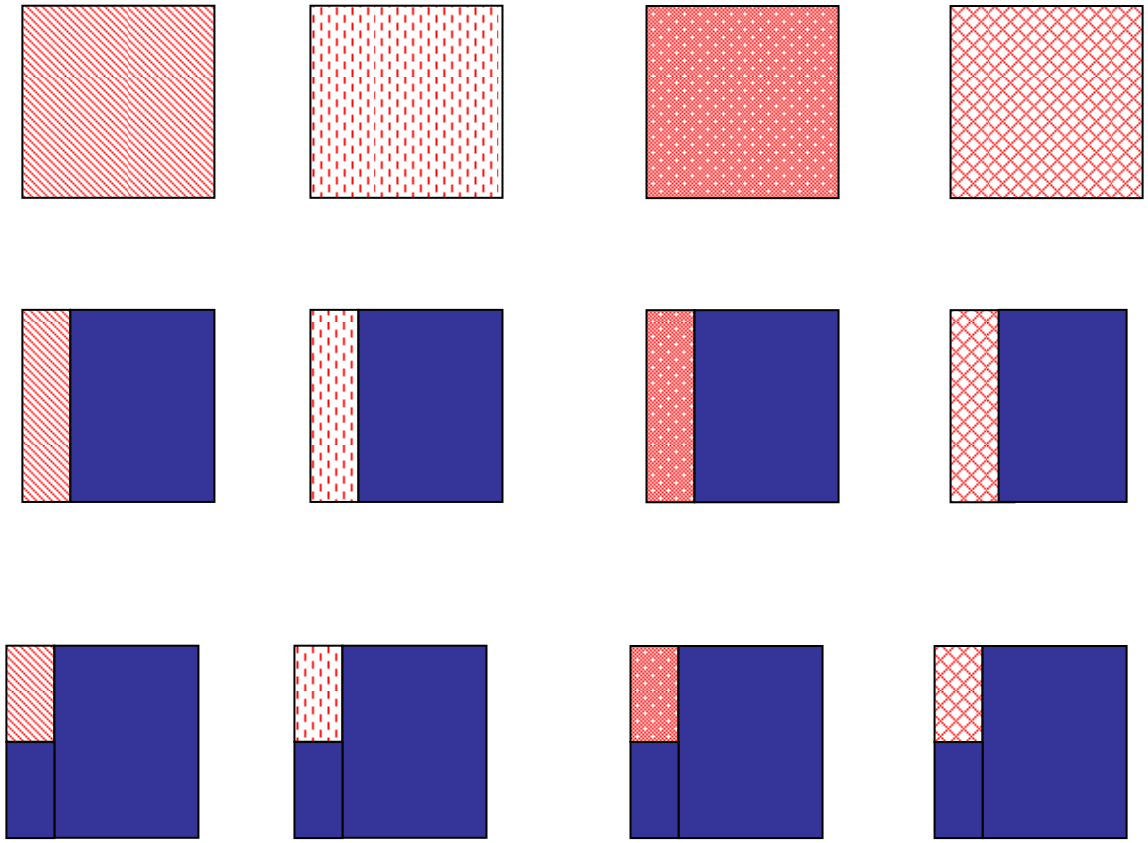


Draws from

$pr(y_{ns} \mid y_s)$

Simple Random sample

Different versions of synthetic samples



We consider full synthesis in this talk given that we want to release geographical details

Setup

Population:

A *areas,* B_i *units in area* $i = 1, 2, ..., A$

$\{\pi_{ij}, i = 1, 2, ..., A;\ j = 1, 2, ..., B_i\}$

Selection Probabilities

Sample:

$\{y_{ij}, i = 1, 2, ..., a;\ j = 1, 2, ..., b_i\}$

Survey Variables

Model

Sampling Mechanism

$[S_{ij} \mid \pi_{ij}, \theta][y_{ij} \mid S_{ij}, \pi_{ij}, \phi]$

Model for outcome

Setup (contd)

$$[b_i \mid B_i, \lambda]$$

Model to determine sample size

$$Synthetic\ Data:$$

$$(\theta, \phi, \lambda) \mid Data$$

$$(b_i^*, S_{ij}^*, y_{ij}^*) \mid \theta, \phi, \lambda, Data$$

$$j = 1, 2, ..., b_i^*; i = 1, 2, ..., a^*$$

Generate multiple synthetic data sets

Example

$$y_{ij} \sim N(\theta_i, \sigma^2), i = 1, 2, ..., a;\ j = 1, 2, ..., b_i$$

$$\theta_i \sim N(\mu, \tau^2)$$

$$b_i \sim Poisson[\exp(\lambda_o + \lambda_1 B_i)]$$

Gibbs sampling can be used to generate synthetic data sets

Are the inferences from the synthetic data sets similar to those from the actual data?

Simulation

$\mu = 0.5, \sigma = \tau = 1$

$\lambda_o = 0, \lambda_1 = 0.19$

$a = 50, A = 100$

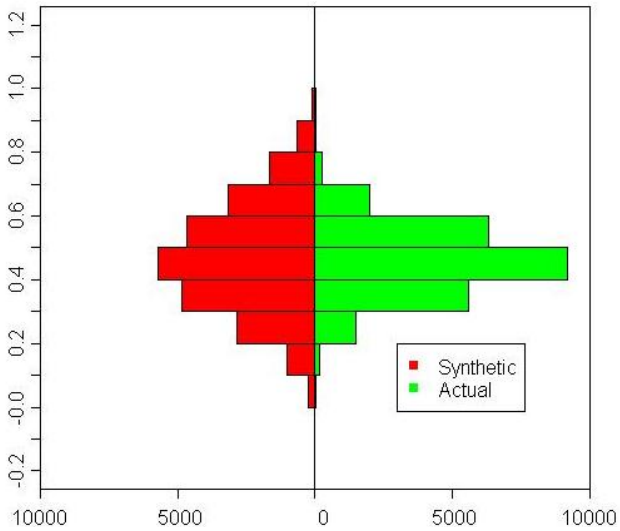
$B_i = (14, 2800), b_i = (5, 527), a^* = A$

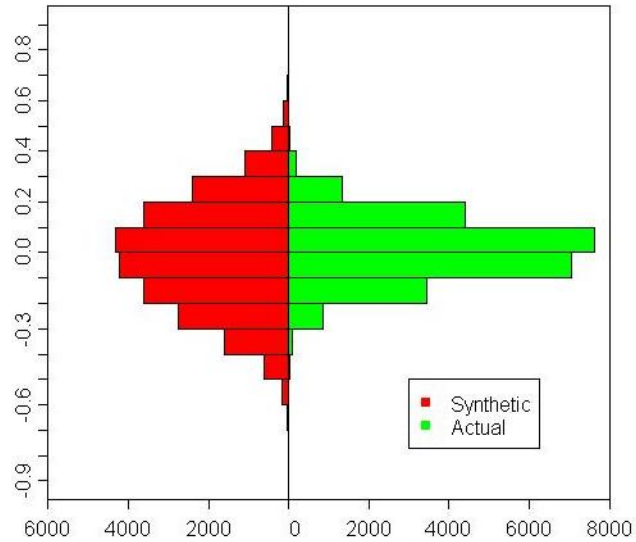
500 replicates

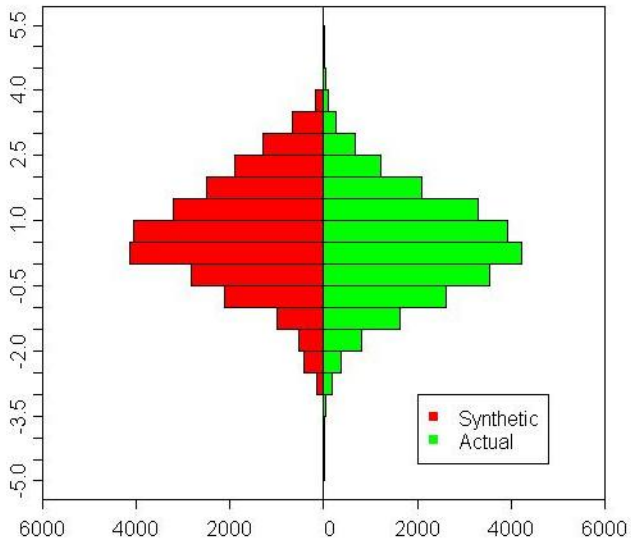
50 synthetic data sets for each replicate

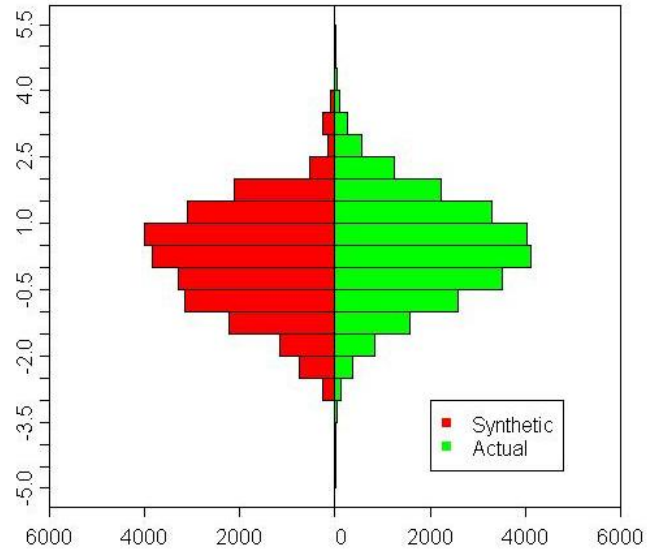
Compared small area estimates, variance and confidence intervals from the observed and synthetic data sets

Posterior distributions of small area means for two sampled (top panel) and two nonsampled (bottom panel) with synthetic data (red) and observed data (green)

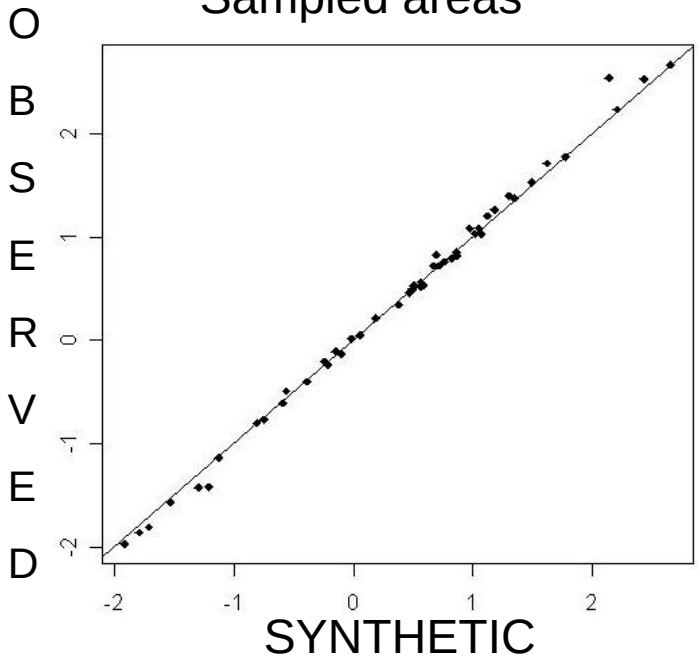




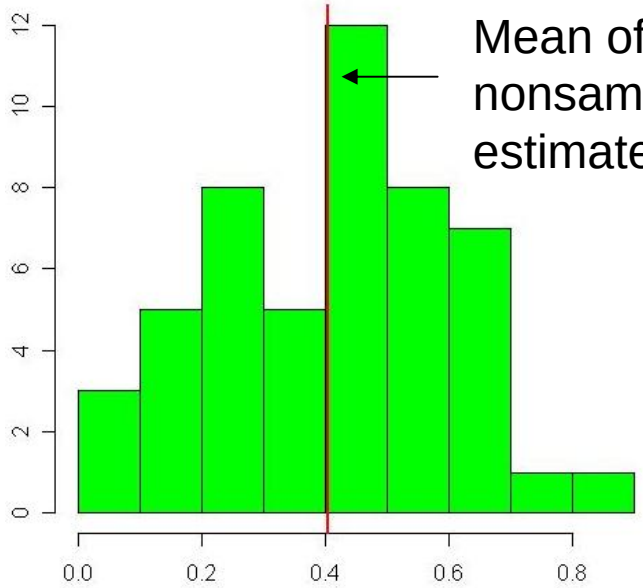




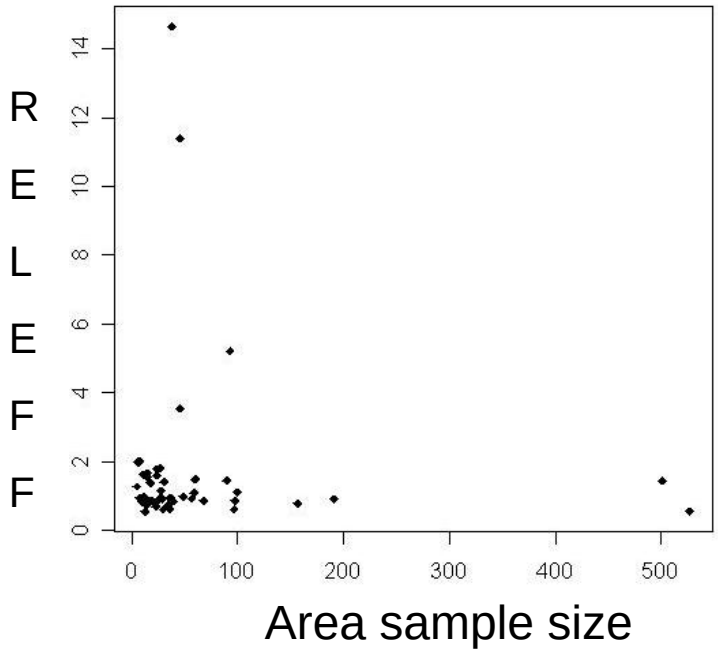
Sampled areas



Non-sampled areas



Relative Efficiency of synthetic estimates compared to observed data estimates as a function of the area specific sample size



Other Measures

- Confidence coverage practically identical for synthetic and observed data sets.
- Probability overlap in confidence intervals ranged from 93% to 95%
- Synthetic confidence intervals were about 17% to 38% wider than the observed data confidence intervals
- Synthetic RMSEs were about 2.8 to 7.1 times larger than the observed data RMSE
- Bottom line:
  - Inference from synthetic data sets were valid from the frequentist point of view.
  - Posterior distributions of area level means similar between synthetic and observed data sets

Future Work

- Consider survey designs with weights
- Consider multivariate outcomes
- Consider nonparametric models for survey outcomes
- Applications to actual survey data. American Community survey is a possibility