

Kinney et al. (2011) Response Questions

Name: Maggie Wu

1. What are the synthesized and un-synthesized variables in the SynLBD?

In the SynLBD, ID, geographic location, and industry code are the un-synthesized variables, whereas first and last years of observed establishment, multiunit status, annual payroll, and annual employment are the synthesized ones.

2. What approaches did they take when they have more than 1 variables deemed sensitive and to be synthesized?

They set an order for synthesizing these sensitive variables and conditioned the synthesis of on the previously synthesized ones. (sequential synthesis & Gibbs Sampler)

3. What Bayesian synthesis model(s) did they use?

For the highly skewed payroll and employment, they transformed them with the kernel density estimation procedure and, with marginally standard normal distributions, they applied the normal model to these variables.

4. How many synthetic datasets did they generate?

They generated one synthetic dataset.

5. How did they evaluate the utility of the synthetic datasets? Can you think of any other utility measures?

They plotted the synthesized data against the original one and added the line of slope 1 to show how much the former aligned with the latter; in addition, they applied this to the regression coefficients to examine if relationships between variables were preserved with their synthesis method.

6. How did they evaluate the disclosure risks? Can you think of any other disclosure risks measures?

They dismissed the identification risks in the data, since “there are no real data on SynLBD,” rendering “little to match with certainty.”