Program in Survey Methodology Dissertation Defense:

Joseph Sakshaug

Chair: T.E. Raghunathan

Synthetic Data for Small Area Estimation

Friday, August 5, 2011

2:00pm

368 ISR and 1208 LeFrak

Small area estimates provide a critical source of information used by a variety of stakeholders to study human conditions and behavior at the local level. Statistical agencies regularly collect microdata from small geographic areas but are prevented from identifying these areas in public-use microdata sets due to disclosure concerns. Alternative data dissemination methods include releasing summary tables for small areas and accessing restricted identifiers via Research Data Centers. This dissertation proposes a new method of disseminating public-use microdata that contains more geographical details than are currently being released. The basic idea is to replace the observed survey values with imputed, or synthetic, values. Data confidentiality is enhanced because no actual values are released. In the first chapter, a hierarchical Bayesian model is proposed to generate synthetic microdata from the posterior predictive distribution. In the second chapter, a nonparametric procedure for generating synthetic data for continuous non-normal distributions is developed. In the third chapter, a synthetic data method that accounts for complex sample design features and permits the generation of synthetic data for both sampled and nonsampled areas is proposed. The methods are demonstrated and evaluated using a mix of public-use and restricted microdata from the American Community Survey and National Health Interview Survey. Each of the methods is evaluated using empirical, simulation, and cross-validation studies. The analytic validity of the methods is assessed by comparing the small area estimates obtained from the synthetic data with those obtained from the observed data.