

Synthetic Longitudinal Business Data

Satkartar K. Kinney*

National Institute of Statistical Sciences, RTP, NC USA saki@niss.org

Jerome P. Reiter

Duke University, Durham, NC USA jerry@stat.duke.edu

Longitudinal business data are widely desired by researchers, but difficult to make available to the public because of confidentiality constraints. In this paper, we discuss the generation of synthetic public use datasets for establishment data. The basic idea is to release simulated values of sensitive variables, generated from probability distributions fit using genuine data. This can protect confidentiality, since attributes are synthetic rather than real. And, when the models describe the data well, broad-scale inferences from the synthetic datasets will be inferentially valid. We discuss the approaches used for generating synthetic public-use files for the U. S. Census Longitudinal Business Database.

Key Words: Disclosure limitation, multiple imputation, establishment data