The Synthetic Longitudinal Business Database: experiences and lessons learned

Javier Miranda
U.S. Census Bureau, Washington, D.C., USA Javier.miranda@census.gov

Lars Vilhuber*
Cornell University, Ithaca, NY, USA lars.vilhuber@cornell.edu

In most countries, statistical agencies do not release establishment-level business microdata because doing so represents too large a risk to establishments' confidentiality. One potential approach for overcoming these risks is to release synthetic data. Here, establishment data are simulated from statistical models designed to mimic the distributions of the real, underlying microdata. The US Census Bureau Center for Economic Studies in collaboration with Duke University, the National Institute of Statistical Sciencies, and Cornell University made available a synthetic public use file for the Longitudinal Business Database (LBD), an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976. The resulting product, dubbed the SynLBD, was released in 2011 and is the first-ever comprehensive business microdata set publicly released in the United States including data on establishments' employment and payroll, birth and death years, and industrial classification. This paper documents the scope of projects that have requested and used the SynLBD. We describe some of the validation results obtained, as well as lessons learned.

Keywords: Synthetic Data, Business Data, Longitudinal Data, United States