LEHD Fact Sheet

Simulation Server for LEHD: Data for State Partners

Purpose
To provide state partners with access to (simulated) micro data from their state's QWI files as well as access to the research tools used at LEHD. Allow original research and development of statistical research that can be applied directly to the confidential data, if needed.

Approach
State partners can access the secure simulation server at Cornell University via web-based login through the Cornell Restricted Access Data Center firewall. They can then log in to the simulation server. LEHD tools for accessing and modeling using the QWI micro data and the QWI aggregate data will immediately be available.

Security

- Each state will have its own work area and access to its own data.
- A state may grant other states access to the data but the default is fully private data areas.
- The simulation server manager will handle account administration.
- The LEHD staff will have access to all areas.

What will be on the site?

- Full, multiple-copy simulation of all QWI micro data files (individual characteristics, employer characteristics, work history, associated files)
- All QWI aggregate data
- LEHD-written programs for using these data

What are the simulated data?
Simulated data are disclosable data with same characteristics as the QWI micro data. The LEHD staff are researching ways to replace confidential data elements with simulated values. Simulated data values are draws from a probability distribution. Only disclosable data and summary statistics are used to define the probability distribution from which simulated values are drawn.

The simulation can be done "outside the firewall" and contain no confidential data. The simulation is done multiple times. This yields multiple copies of the database (implicates), each with unique simulated data values in place of confidential data elements. Analyses can be performed separately on each implicate using traditional methods (no special tools or techniques required). Statistics estimated on each implicate are combined using standard methods for multiply-imputed data.

- Variation in the estimates obtained on different implicates is an indicator of the reliability of the estimates.
- Original analyses can be conducted using the remote access data without bring the application directly to Census/LEHD confidential data.
- Custom analyses developed at the remote site can be ported directly to the LEHD QWI system.

Supporting documentation
See "Disclosure limitation in longitudinal linked data" Abowd and Woodcock (2001) in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies edited by P. Doyle, J. Lane, J Theeuwes and L. Zayatz, North Holland, Amsterdam, 2001.
(http://www.elsevier.nl/inca/publications/store/6/2/2/1/2/9/index.htt)