INFO 7470/ILRLE 7400
LEHD RDC Data Products

Kevin McKinney
U.S. Census Bureau
and
Lars Vilhuber
Cornell University

March 8, 2011

LEHD Infrastructure in the Census Research Data Centers

- Big-picture: http://www.vrdc.cornell.edu/news/lehd-infrastructure-files-in-the-census-rdc-overview/ (overview_master_zero_obs.pdf)
- Contains
  - detailed file descriptions
  - Attached zero-obs versions of all datasets

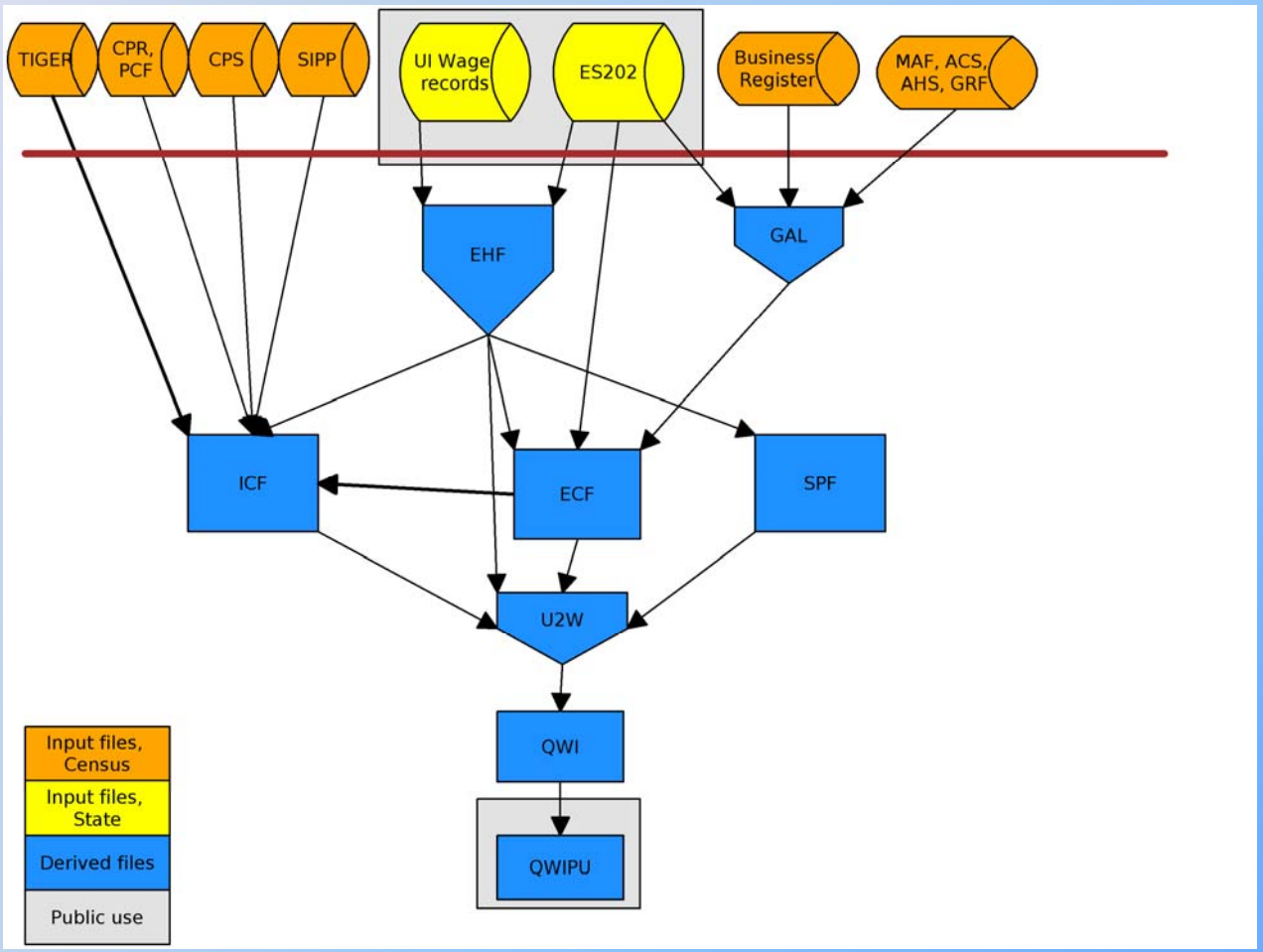
Important features of RDC LEHD data

- No public-use QWI information (non-fuzzed data only)
- no information related to the disclosure-avoidance measures used in QWI and OTM
- Certain lag relative to the publicly available QWI
  - Currently: Snapshot 2004
  - Ready, but not yet available: Snapshot 2008
- Not all states have given permission to use their data for non-core research purposes

Treatment of Federal Tax Information

- Some components have Title-26 protected variables
- Policy office requires IRS approval for all LEHD projects.
- The LEHD Snapshot isolates T26 data, which in the future may allow projects that do not need IRS approval

Data flow view of LEHD Infrastructure



Name	T26 component	Notes
Business Register Bridge (BRB)	(all)	
Employer Characteristics File (ECF)	ECFT26	CA EIN Also: do not contain firm names
ES-202 (QCEW)	Not Available to researchers, but may be feasible for special projects	
Individual Characteristics File (ICF)	ICFT26	IRS 1040 Residence Addresses S2004: only 1999 address S2008: planned longitudinal addresses
Geocoded Address List (GAL)	GALT26	BR Records

Name	T26 component	Notes
Quarterly Workforce Indicators (QWI, establishment level files)		Unfuzzed
Successor-Predecessor Files (SPF)		Preliminary access (documentation is beta quality)
Unit-to-Worker Impute (U2W)		Multiply-imputed files, should use multiple imputation analysis techniques

Determining Data Availability

- Use overview_master_zero_obs.pdf
- Table 1.2 – States for which ANY files are available (permissions given)
- Table 1.3 – Files that exist for each state (EHF, ECF, and ICF are the core processes)
- Cross-reference Table 1.2 with Table 1.3
- Each process has a table with the available time periods. For example, see table 4.7 on page 127 for the EHF.

Identifiers

- In general, linkages between the different files occur using deterministic match-merge techniques
- Person, firm, and establishment identifiers link all LEHD Infrastructure files among themselves.
- External linkages are generally probabilistic:
  - Linkages to BR (many-to-many match)
  - Linkages to external files by establishment location
  - Linkages to external files by firm name (special projects only)

Individual identifier system (PIK)

- All Social Security Numbers (SSN) have been replaced by Protected Identity Key (PIK)
  - no SSN's are available anywhere in this data.
- A PIK is a unique 9 digit number that maps to one and only one SSN.
- A PIK is permanently assigned to an SSN, allowing the same types of analyses to be performed, but with greatly improved confidentiality protection.
- Used widely for person-level files within the Census RDC

Additional identifiers

- Survey IDs
  - CPS
  - SIPP
  - Census 2000
  - ACS
- In general, PIK is on the file, or available as a separate crosswalk
- *Note: LEHD ICF contains CPS and SIPP IDs, but they are not the most current – use the crosswalks*

Firm/establishment identifiers

- Firm identifiers are called *State employer identification number* (SEIN) and generally reflect an entity reporting Unemployment insurance data (UI) taxes to state authorities.
- “Establishments” (more precisely: reporting units) are identified by a combination of SEIN and reporting unit (SEINUNIT).
- The firm and establishment identifiers are state-specific - within the LEHD Infrastructure, there is no method of linking units of a nation-wide firm across state borders.
- Federal EIN is available
  - on ECF for most states,
  - on ECFT26 for California
- CFN availble on BRB

EHF

- PIK SEIN SEINUNIT YEAR files
  - But: Only Minnesota currently has SEINUNIT (establishment) identifiers.
- Contains person-firm (job) x year information.
  - Quarterly earning records
- No direct measure of labor force attachment, it is inferred by the presence of earnings.

EHF auxiliary files

- Auxiliary PHF (person-history file): PIK-SEIN-SEINUNIT
  - Wide file, entire job history arrayed out
  - Employment indicator flag (character string, 0001111000…)
  - Earnings for each quarter
- Auxiliary Unit History File (UHF)
  - SEIN-SEINUNIT history of establishment activity, based on ES202
- SEIN History File (SHF)
  - SEIN history of firm activity, based on ES202
- List of unique PIKs (ever appeared in a state)
- Controltotals: BLS public-use employment (private only) for equivalent time period, can be used for weighting

ICF

- PIK level file
- Contains person level demographic information, primarily from SSA applications for a SSN.
  - Sex
  - Race
  - DOB
  - POB
  - Education
- Completed files: missing characteristics have been imputed (flagged, and multiple imputes available in auxiliary files)
  - Impute rate particularly high for Education
- Recently, higher quality imputations, especially important for education, have been developed. In use at LEHD, but will only be present in the next Snapshot (S2011 maybe)

ICF imputations

- Completed files: missing characteristics have been imputed (flagged, and multiple imputes available in auxiliary files)
  - Impute rate particularly high for Education
- Recently, higher quality imputations, especially important for education, have been developed.
  - In use at LEHD, but will only be present in the next Snapshot (S2011 maybe)

ICF auxiliary files

- Auxiliary files contain array of multiple imputes for those PIKs that had some missing data
  - icf_zz_implicates_age_sex.sas7bdat
  - icf_zz_implicates_county.sas7bdat
  - icf_zz_implicates_education.sas7bdat

ECF

- SEIN SEINUNIT YEAR QUARTER files and SEIN YEAR QUARTER files.
- Contains establishment and firm level information.
  - Location
  - Industry
  - Firm Size (employment and payroll)
  - Public / Private
- QCEW-derived point-in-time measures of employment are available for the 12th of every month.

ECF imputations

- Completed file: All missing information is imputed, through variety of methods
  - NAICS/SIC through bi-directional empirical (probabilistic) crosswalks
  - Employment measures based on alternate employment measures (previous quarter, UI <-> QCEW measures)
  - Establishment location based on distribution of similar establishments within state
- Contrary to most LEHD files, no multiple imputation
- Abundance of flags to undo most imputations

QWI establishment level

- Establishment level (YEAR-QUARTER-SEIN-SEINUNIT) tabulations
- Same statistics as public-use QWI (accessions, separations, etc., by demography), but at the establishment level
- Already incorporate U2W multiple records

Addresses: GAL

- GAL is a list of all unduplicated addresses from
  - Business Register
  - ACS (place of work)
  - AHS
  - ES-202
  - Census Master Address File (MAF)
- All establishment addresses from LEHD Infrastructure have been geocoded to the most accurate level possible
- GAL does not contain residential addresses (see ICF T26 components)
- Cross-walk to the ECF is available

Internal consistency of LEHD Infrastructure

- LEHD Infrastructure is constructed to be internally consistent
  - Firms on EHF = Firms on ECF (superset of UI and ES202)
  - Individuals on EHF = Individuals on ICF
- … and complete: all missing data is imputed or edited
  - Addresses (generally to block levels, at least to county level)
  - Periodically missing information on firms
  - (research) periodically missing information on individuals/jobs, due to firm or state-level non-reporting.

Working with files

- LEHD Infrastructure files are huge when compared to regular research files. In the (S2004) version, in all existing states and years combined, there are
  - 6,100,912,201 wage records
  - 754,775,697 unique jobs
  - 226,639,116 quarterly observations on firms.
  - Total size of all datasets is about 1.5TB
- Careful planning is required to ensure that adequate resources are available.
- Careful programming is required to make analysis feasible.

Working with files

- Random variables can be used to select a sub-sample of persons, establishments, or firms.
  - ECF SEIN level: `sample_sein`
  - ECF SEINUNIT level: `sample_seinunit`
  - ICF: `substr(PIK,1,2)`
- No equivalent variable provided for EHF (jobs)
- Industry or other person / firm characteristics can also be used to create a smaller analysis dataset.

Sample program: on-the-fly sampling

```
%let state=tx;
libname INLIB
  "/mixedtmp/lehd/s2004/ecf/&state./";
data mydata/view=mydata;
   set INLIB.ecf_&state._seinunit
   (where=(sample_seinunit <= 0.05));
run;
proc reg data=mydata;
  model y= x w z;
run;
```

Common errors

- Administrative (SEIN or EIN) firms are not the same as an economic firm.
  - Firms are dynamic entities.
  - Single versus Multi-unit firms.
  - SEIN Firms are *state* based entities.
- EHF earnings are at the SEIN (firm) level, while the ECF contains data at the establishment level.
  - U2W can be used to assign workers to an estab using a multiply-imputed assignment
- Multi-unit reporting is required, but there is no penalty for not breaking out establishments.

Warnings

- BRB bridge links CES Economic data to LEHD at the EIN STATE COUNTY SIC2 level. Multi-unit establishments may not link one to one.
- Current ***education*** variable is of very low quality.
- ECF (LEG) and GAL are not internally consistent on S2004 version of the snapshot (corrected in S2008).
- GAL users must handle missing location information themselves.

Special Disclosure Avoidance Rules

- In general, only model output is allowed
- Any tabulations are subject to QWI rules
  - National, and aggregates of states are probably acceptable, but may require review by the Census Disclosure Review Board (DRB).