Understanding LEHD and SYNTHETIC HOME TO WORK FLOWS IN "ON THE MAP"

DRAFT May 4, 2007

Elaine Murakami, FHWA Office of Planning

Elaine.murakami@fhwa.dot.gov

206-220-4460

This document provides a simple explanation of Longitudinal Employment and Household Dynamics (LEHD) for transportation and economic development, and explains LEHD's future potential, highlighting numerous issues. LEHD is a project of the U.S. Census Bureau that uses the Bureau of Labor Statistics (BLS) Quarterly Census of Employment and Wages (QCEW), Unemployment Insurance (UI) files, combined with federal administrative records. "On the Map" is one component of the LEHD that synthesizes home-to-work flows at the Census block level. This document should be reviewed in conjunction with many of the more technical papers explaining the LEHD process. It is intended for planners, state labor market information department personnel, and state-level policymakers.

Abbreviations

AASHTO	American Association of State Highway and Transportation Officials
ACS	American Community Survey
CTPP	Census Transportation Planning Products
DOT	Department of Transportation
FHWA	Federal Highway Administration
LEHD	Longitudinal Employer-Household Dynamics
MWR	Multiple Worksite Report
NAICS	North American Industrial Classification System
PIK	Protected Identification Key
QCEW	Quarterly Census of Employment and Wages
SIC	Standard Industrial Classification
SSN	Social Security Number
StARS	Statistical Administration Records System
TAZ	Transportation (or Traffic) Analysis Zone
UI	Unemployment Insurance

WHY ARE THE TRANSPORTATION COMMUNITY AND ECONOMIC COMMUNITY INTERESTED IN LEHD AND "ON THE MAP"?

Transportation planners have used decennial Census "long form" data for home-to-work data for over 35 years. Since 1970, the transportation community has paid for a special tabulation of the long form data that produced small area home-to-work flow tabulations. Since 1990, there has been a nationwide program sponsored under the auspices of AASHTO (American Association of State Highway and Transportation Officials) to

assure that all MPOs and each state had the data readily available for transportation planning purposes. Over the years the special tabulation effort has gone through several changes and is currently known in the industry as the Census Transportation Planning Products (CTPP) program. For a detail history of the special tabulation program refer to http://www.trbcensus.com/articles/ctpphistory.pdf.

For the CTPP2000, small area flows included Traffic Analysis Zone (TAZ)-to-TAZ and Block Group-to-Block Group flows for some areas. However, the decennial census "long form" has been eliminated for 2010, and replaced with the American Community Survey (ACS). While the transportation community is planning to use a five-year accumulation of ACS records to create several CTPP products, the LEHD "on the map" is a potential alternate source for home-to-work flow data. The LEHD offers to provide updated data as frequently as every two years or even every quarter.

Block-to-Block LEHD transportation origination/destination data can be used in journey-to-work analysis, both for labor sheds, where residents reside and who commute to a given place, or commute sheds (where workers work). Such data can improve travel demand forecasting, improve the ability to more precisely target prospective customers for ridesharing and commute trip reduction marketing, improve bus route planning and aid in jobs/housing analysis. LEHD origination/destination data can also contribute to sophisticated real-estate location decision analysis, allowing comparison of available labor force between different places.

FHWA has been encouraging State DOTs and Metropolitan Planning Organizations (MPOs) to work together with their State Employment Security Departments and Labor Market Initiatives office to obtain a copy of the "On The Map" data, and to examine the data to understand the benefits of the data, and to recommend improvements, particularly in the area of feedback mechanisms to improve business address information.

The first release of "On the Map" data used 2002 and 2003 QCEW data, and was initially available for 12 States. LIMITED examination of the "On the Map" data has identified several problems.

- School districts often linked to district headquarters office, rather than a specific school. The district headquarters' address is incorrect. (Colorado)
- State employees most often linked to an office in the State Capitol. (Illinois)
- Too many "out of state" commutes were identified, for example 40% of workers from "out of state" in the "On the Map" data, compared to 2% from "out of state" using the CTPP2000. (Illinois) Note: algorithm for 2004 data has been changed and CB says the results are comparable to CTPP2000. For example, 6 percent of workers in Cook County from out of state.
- Average distance of home-to-work exceeding 25 miles where CTPP2000 found about 12 miles (more consistent with regional survey and model results). (California)

We do not believe that these problems are isolated incidents, but instead may reflect systematic problems that have not yet been identified and resolved. We cannot say whether these problems are specific to one or many states, or specific to certain industries. This is not to say that these problems occur across the board, but means that the data should be REVIEWED thoroughly BEFORE it is used for transportation planning applications.

Data users need a better understanding of the underlying data and the data synthesis process in creating the home-to-work flow data presented in "On the Map."

WHAT IS THE UNIVERSE OF WORKERS IN LEHD ON THE MAP?

The QCEW records include "covered" employment, that is, workers who are covered by unemployment insurance. This EXCLUDES self-employed, railroad, and federal government workers. Nationwide, about 10 percent of workers are self-employed, and less than 1 percent are federal government workers. It also excludes people in the "informal economy," people working for cash without any administrative records for wages or taxes. Graham and Ong estimates that "informal" workers could be as much as 14 percent of total workers in the Los Angeles metropolitan area. [1]

The QCEW records include multiple employers for individual workers. Multiple employers for an individual can occur for many reasons including: a person got a new job during the year; a person has multiple jobs at any given time, e.g. two different part-time jobs. In "On the Map," a "primary" job is selected by selecting the job with the highest wages. The use of administrative records collected over one year in OTM differs significantly from the decennial Census and the American Community Survey approach where the survey is conducted at a specific point in time, and only one work location can be answered on the survey form.

The spatial accuracy of the QCEW data varies widely. The States only have to assure the accuracy of data to the County level. The amount of effort each State has expended to improve spatial accuracy, by checking addresses, identifying payroll processing offices that differ from work locations, and checking completeness of multiple worksite addresses varies.

WHERE DO THE WORKPLACE LOCATIONS COME FROM?

Step 1. Establish a list of potential workplace locations.

Workplace addresses come from the QCEW. Business "firms" are required to report a least ONE location per State for the QCEW record if they have paid workers in that State. As part of QCEW there is a Multiple Worksite Report (MWR). As of 2006, about 50 percent of States now have State laws REQUIRING completion of the Multiple Worksite Report. Each year, about 1 or 2 more states has shifted to mandatory reporting of

[1] Graham, M. R. and Ong, P. Social, "Economic, Spatial and Commuting Patterns of Informal Jobholders" Technical Paper No. TP-2007-02 http://lehd.dsd.census.gov/led/library/techpapers/tp-2007-02.pdf

multiple worksites. This report includes the address location and the number of employees (per month) for each worksite with the total quarterly earnings for that site. (per Dave Higgins at BLS 202-691-6460) http://www.bls.gov/cew/cewmwr00.htm

For LEHD, the QCEW is supplemented with records from the Census Bureau's Business Register (BR) using the Business Register Bridge (BRB). http://lehd.dsd.census.gov/led/library/tech_user_guides/brb_ces_master.pdf I have not found documentation on the scale of additions, types of business addresses (e.g. by different industries) that are added to the list of businesses using the BRB. We have not done research to determine the completeness of the BR, particularly for businesses with multiple sites.

Known Issues in Step 1:

- Large employers vary widely in their compliance with completing MWR.
  - Firms with multiple worksites are listed as having only one worksite.[2]
  - Firms identified with multiple worksites, have only an incomplete list of sites
- The level of effort by State Employment offices to obtain Multiple Worksite Reports for large employers varies widely. This is true for states both for which MWR are mandatory and for which MWR are voluntary.
- Workers employed by a Professional Employment Offices (PEOs) are geocoded to the PEO address, rather than a location where they are working day to day. PEO's, sometimes called "personnel supply companies) provide staffing services to businesses to handle payroll, timecards, benefits. This is a known problem in Florida, but may be increasing in other states. For example, USDOT uses many IT contractors who sit at the USDOT headquarters office in Washington, D.C. However, their employer is located somewhere else. A few states now require PEOs to report individual addresses for the location of where the work is being conducted.

---

[2] Julia Lane and Marc Roemer, Mix, Putnam, Almousa, O'Connell and Foster. ""An Evaluation of the Use of the LEHD Data for Transportation Planning" Final report. September 16, 2003.

p. 11 includes an itemization of large employers with a single record.

"Florida .one file contained 17 establishments for employers that have greater than 10,000 employees but only one unit." These 17 establishments accounted for 357,486 workers. The other four files contained information on employers with 1,000 to 10,000 employees and only one unit." Breakdowns by industry are included, e.g. 49 medical establishments, 45 education establishments.

"Illinois.one file contained 7 establishments for employers that have greater than 10,000 employees but only one unit." These 7 establishments accounted for 157,323 workers. A similar breakdown for establishments with 1,000 to 10,000 workers is provided.

- State and Local Government workers are often assigned to one location only (similar to business establishments). USDOT Bureau of Transportation Statistics (BTS) sponsored research found that in IL, state government workers were nearly always assigned to Springfield, IL.
- School employees are often assigned to a School District office, rather than to individual schools. This could probably be corrected using the National Center for Education Statistics/Common Core of Data file which include a physical address, the number of teachers, and the school district name. http://nces.ed.gov/ccd/search.asp
- Some addresses are payroll processing offices and do not reflect work locations.

Step 2. Assign workers to a specific worksite using a model based on MN data

The MWR does not link individual employees by SSN to a specific worksite. For example, if a grocery store chain lists 10 different stores in the MWR, an individual SSN is not linked to a specific store location. Only in Minnesota (MN) does the Unemployment Insurance (UI) file with SSN provide a link to a specific worksite. Therefore, to assign individual workers to a specific worksite in all other states, a model was built using data from MN. The individual worker's home address is used in combination with the worksite location to calculate distance from home.

This model uses several variables including:

- Employee characteristics such as distance from home, and length of employment at that firm.
- Establishment characteristics such as size of the firm, and "age" of the firm.

Step 3. Add noise to protect confidentiality of firms

The total number of employment (and earnings) will not exactly match confidential data from each State because of the addition of noise. However, the amount of noise is small and should not result in overall distortions to the data

WHERE DOES RESIDENCE LOCATION COME FROM?

Residence address of workers is taken from the Statistical Administrative Research System (StARS). StARS combines several federal administrative files, such as Social Security, IRS, Medicare, Medicaid, Veteran's Affairs, that include SSN. The residence address in StARS is coded to a census block (or block face?)

Both StARS and the Unemployment Insurance (UI) record include SSN. SSN is translated into Protected Identification Key (PIK), using an algorithm that is held constant so that as new UI files are available, the same PIK is created for historical matching (and analysis of work force dynamics). The UI file is joined to the StARS file by PIK to provide a residence location.

Situations where addresses in StARS may not be where the person is residing

- College/university students are most likely to report a parent's address for their W-2, and Social Security records.
- Summer (or other temporary) employment may also create problems where a permanent address is used rather than an actual residence location.
- Potential problems with incorrect SSN linkages due to undocumented workers using "borrowed" SSNs.
- Moving residences within a year. Most residence information is taken from IRS 1040's (April 1), but primary job is selected based on highest earnings.

DISCLOSURE PROOFING THE DATA BY CREATING SYNTHETIC FLOWS

To protect the confidentiality of individual workers, the home-to-work flow is synthesized. The counts of workers and jobs in a specific workplace block are "real" (with some added noise (see above). However, the distributions between the residence block and the workblock are synthetic.

The synthetic place of residence distributions are created for each unique combination of 4 attribute variables of work block. These 4 attributes result in a matrix of 360 cells (20 x 2 x 3 x 3 = 360).

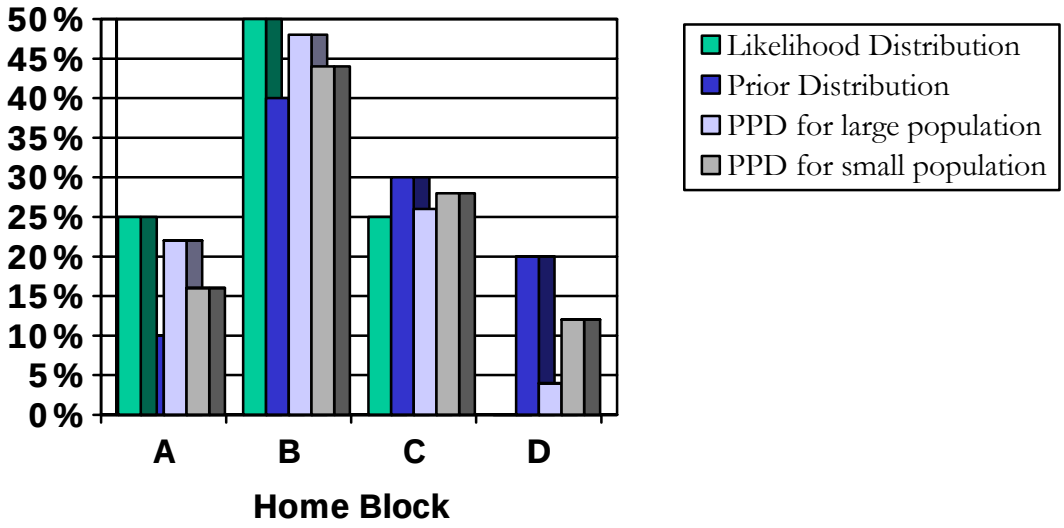
Firm attributes:

- industry (20 NAICS[3] sectors),
- firm ownership (2 categories: private or public (government and non-profit))

- Worker attributes:
  - Age (3 categories) (0-30, 31-54, and 55 and over), and
  - Earnings (3 categories : 0-$1,199; $1,200-$3,399, $3,400 and over

The creation of synthetic data is illustrated in the graph below, developed by Fredrik Andersson from the Census Bureau. Although it shows only 4 choices for home block (A, B, C, and D), the model requires at least 10 potential residence blocks to be used for each cell (or aggregations of cells). Ten implicates of synthetic data are created.

[3] NAICS is the North American Industrial Classification System. It replaced the Standard Industrial Classification system (SIC).



The green bars show the actual place of residence distribution. 50% of the workers reside in block B, and 25% in each of block A and C, and none in block D. To protect individual confidentiality, a synthetic distribution, called a Prior, is created, shown as the Blue bars. The Prior is created by using aggregated values of the 4 attributes listed above (industry sector, ownership, age, and earnings). For example, add residence blocks to the choice set using workers from the same work location and industry, but from workers who do not match the age or earnings categories. The green bars are combined with the blue bars to create a Posterior Predictive Distribution (PPD), with different distributions depending on the number of workers at the workplace location.

As the number of workers at the work place increases, the assignment of a worker to an actual residence block from workers at the same workplace increases. If the number of workers at a given workplace is small, the greater the chance that the residence block may be chosen from a larger group of residence blocks, some "real" and some synthetic.

RECENT FINDINGS FROM LOCAL REVIEW OF 2002/2003 ON THE MAP

Recent comparisons using CTPP2000 and LEHD for the San Francisco Bay Area, and Skokie, IL raise questions about the accuracy of the LEHD On the Map data.

San Francisco Bay Area Hacienda Business Park

	CTPP 2000	LEHD On the Map
Within 5 miles	25%	10%
Within 10 miles	40%	21%
Within 20 miles	65%	38%

Source: Steve Raney, Cities 21 http://www.cities21.org/BABPC/

A team of researchers (Soot, Metaxatos, Thakuriah) at the University of Illinois, Chicago (UIC) identified two key issues with the LEHD On the Map data. First, they found several work locations where very high proportions of workers travel to work from out-of-state. Several neighborhoods (suburbs of Chicago) were found to have 30 to 40 percent of workers from out-of-state, where it would be expected that less than 5 percent would have these long commutes. For two jurisdictions, Cicero and Winnetka, we extracted the CTPP2000 results, and found less than 2 percent of workers commuting from out-of-state. A check of 1990 CTPP showed similar results. The Census Bureau has revised their algorithm for out-of-state commuting in the new 2004 OTM data release, and says their results are much more comparable to the CTPP2000 results. For example, they are finding that 6 percent of workers in Cook County, IL coming from out-of-state.

Second, after removing the "out of state" pairs, and re-weighting the remaining home-to-work pairs, the LEHD On the Map showed a similar pattern to the SF Bay Hacienda Business Park, where a much higher proportion of workers had trips longer than 25 miles. For example, for people working in Skokie, IL, about 13 percent with trips longer than 25 miles, compared to 4 percent using CTPP. Similarly, for people who live in Skokie, the On the Map resulted in about 12 percent with trips longer than 20 miles, compared to less than 3 percent in CTPP2000.

CONCLUSIONS

We highly recommend that transportation planners treat the LEHD On the Map 2004 dataset as an exploratory research dataset. The data should be reviewed and evaluated. It should be validated against other data sources. Until a thorough review is completed, users should use the data with caution. Problems should be conveyed to the LEHD project team, so that problems can be researched and resolved.

We believe that in the long run, OnTheMap (OTM) will prove to be a valuable tool for transportation planners. We want to encourage evaluation by the transportation data community and want to encourage the transportation data community to work with State employment offices to establish better feedback loops to improve the data quality. We believe that the potential utility of OnTheMap is high, but that the Census Bureau must educate users about the data sources and the data synthesis process before marketing it widely as a tool for transportation applications.

Acknowledgements:

Thank you to Fredrik Andersson of the U.S. Census Bureau for providing the PPD bar chart and explaining synthetic flows. Thank you to David Higgins at BLS for providing information about Multiple Worksite Reports. Thank you to Steve Raney at Cities21 for recommending that a short document on the LEHD On the Map process be written.