REGION
The Journal of ERSA
Powered by WU

Volume 5, Number 3, 2018, R5–R12
DOI: 10.18335/region.v5i3.251

journal homepage: region.ersa.org

The US Census Longitudinal Employer-Household Dynamics Datasets

Robert Manduca[1]

[1] Harvard University, Cambridge, MA, USA

Received: 2 August 2018/Accepted: 6 August 2018

Abstract. Over the last several years, the Longitudinal Employer-Household Dynamics program at the US Census Bureau has partnered with state labor market information offices to produce a collection of extremely rich datasets based on linked employer-employee records. These datasets, available free for download from the program website, offer exceptionally detailed information on a number of topics of interest to regional scientists, including migration, local labor market dynamics, and the spatial distribution of employment. This article describes the different publicly available datasets, the process by which these data are generated, and examples of research in regional science that is already using these data.

1 Introduction

Regional science is becoming an increasingly data-intensive field. More and more, scholars require highly detailed and comprehensive information about regional economies in order to accurately characterize them. This change requires moving beyond the traditional data sources that regional scientists have used, and incorporating new datasets, often compiled from administrative records.

One new data source of interest to regional scientists studying the United States (US) is the data produced by the Longitudinal Employer-Household Dynamics (LEHD) program at the US Census Bureau. Based on a unique collaboration between the Census Bureau and labor market information offices in all 50 states, the LEHD provides scholars and the public with free access to highly detailed information produced from linked employer-employee insurance records at an unprecedented spatial resolution.

2 The LEHD Infrastructure Files and main data products

At the core of the LEHD data project are the LEHD Infrastructure Files. These confidential files are constructed by linking state unemployment insurance records with the Quarterly Census of Employment and Wages, the Decennial Census, and other administrative and survey datasets. The result is a set of longitudinal files that track individuals who are covered by unemployment insurance over time. The Infrastructure Files themselves are confidential, but available at Federal Statistical Research Data Centers to scholars with proposals accepted by the Census Bureau. The Infrastructure Files have been used extensively in scholarly papers within labor economics and to study topics of interest to regional scientists such as the spatial mismatch hypothesis (Andersson et al. 2018) as well as trends in migration (Hyatt et al. 2016).

While the Infrastructure Files offer an extremely powerful tool for studying labor market dynamics at all spatial scales, they may be impractical to access for many researchers, particularly those based in Europe. To provide easier access to much of the information collected through the LEHD program, the program offers three publicly available datasets downloadable from its website. These are the Quarterly Workforce Indicators (QWI), the LEHD Origin-Destination Employment Statistics (LODES), and the Job-to-Job Flows dataset (J2J).

2.1 Quarterly Workforce Indicators

The QWI is the flagship data product produced by the LEHD program. The QWI provides quarterly (i.e. four times per year) employment and earnings information at the county level. The power of the QWI data comes from the ability to cross tabulate these statistics by a variety of industry, firm, and employee characteristics. The data are disaggregated into 4-digit NAICS Industry Groups, a detailed classification with categories such as "Communications Equipment Manufacturing" and "Vegetable and Melon Farming." The data can be further disaggregated by firm characteristics, such as firm age and firm size, and by employee demographic characteristics, including education level, gender, and race. With the QWI data, it is possible to answer question such as "are women and men working in Silicon Valley web startups paid similar amounts?" or "is the manufacturing decline in Cleveland affecting workers of all races similarly?" More comprehensively, one could ask, "which counties have the largest gender gaps in tech?" or "where are manufacturing startups located?"

The total time span covered by the QWI data varies based on when each state joined the program. Data for all states is available from Q1 2010 to Q2 2016. For many states, including California, Oregon, Washington, Illinois, North Carolina, and Maryland, data extends back until the early 1990s. The available date ranges for each state are posted at https://qwiexplorer.ces.census.gov/loading_status.html.

2.2 LEHD Origin-Destination Employment Statistics

The LODES data compliment the QWI. Whereas the QWI data provide extremely detailed job characteristics at a relatively large level of geography, the LODES data provide limited job characteristics at very fine geographic resolution. Data are provided by census block, the smallest nationally-defined spatial unit, which correspond to city blocks in size. The LODES data are provided state by state in three files. The Workplace Area Characteristics (WAC) file provides information on the jobs located in each block. In addition to the total count of jobs located in that block, the WAC file gives the count of jobs in each of the 20 2-digit NAICS sectors, in each of three income categories, and by employee race, ethnicity, age, sex, and education level. This dataset is perhaps the most comprehensive and detailed publicly available data on the spatial location of employment throughout the US.

The Residential Area Characteristics (RAC) file mirrors the WAC file, but provides characteristics for the workers who live in each block rather than those who work there. The RAC file can be used to determine the rough industry or education breakdown of the residents living in a particular census block.

Finally, the Origin-Destination (OD) file provides information on commuting patterns. For each pair of census blocks, the OD file gives the total count of workers who commute between them, i.e. a person that lives in census block A and works in census block B. The OD file also provides limited information on the characteristics of both the workers and their jobs, including three industry categories, three income categories, and three worker age categories. The OD file is similar in purpose to the data in the Census Transportation Planning Package (CTPP), and can be analyzed similarly. The main difference is that the LODES OD file contains less demographic information but a much higher sample size and much finer spatial resolution. Compared to the CTPP, the LODES data show longer commutes and a smaller fraction of within-county commutes, likely in part because of differences in the sampling frame and in part because of the difficulties in assigning

workers in some multi-establishment employers to particular establishments described below (Graham et al. 2014, Green et al. 2017).

Each of these files is provided for all states and all available years. Note that certain state-year combinations are not available, though data for all 50 states are available for the years 2011-2013. Separate versions of the three files are available for all jobs, private sector jobs only, and federal government jobs only. "Primary job" versions of the files are also available that include only one job per worker, the job from which they earned the most money during the second quarter of each year. The baseline specification for most analyses will be all primary jobs.

In addition to downloading the raw data files, scholars and analysts can interactively extract data and conduct analysis on custom areas using the OnTheMap tool at https://onthemap.ces.census.gov.

2.3 Job-to-Job Flows

The third major dataset available from the LEHD program is the Job-to-Job Flows (J2J). This innovative dataset, still in its beta version, provides information on how people transition between jobs as well as between employment and nonemployment. A third to half of all hires in the United States involve movements between employers rather than from nonemployment (US Census Bureau 2017). These movements between employers are important for job-employee matching – on average, such movements involve an 8% increase in earnings – and are particularly sensitive to the economic climate (Fallick et al. 2012). The J2J data allow detailed study of job transitions, including topics of great interest to regional scientists such as the relationship between job change and migration.

The J2J data consist of two types of files. The main files provide tabulations of hires and separations to other jobs and to nonemployment by geography (state or Metropolitan Statistical Area), firm demographic characteristics (size and age), NAICS 2-digit sector, and worker demographic characteristics. These are provided as both raw counts and as rates, and in both raw and seasonally adjusted form (because hiring has strong seasonal patterns, many analyses will benefit from seasonal adjustment). The second set of files consist of origin-destination data, and report flows of workers moving between industry-geography combinations by firm and worker demographics. Further details about the J2J data and the data creation process are provided in the technical paper "Job-to-Job Flows: New Statistics on Worker Reallocation and Job Turnover" (Hyatt et al. 2017).

J2J data are available quarterly for most states from Q2 2000 until Q1 2018. The data can also be accessed interactively through the Job-to-Job Flows Explorer at https://j2jexplorer.ces.census.gov.

3 Creation of the LEHD files

The LEHD dataset is created through a partnership between the Census Bureau and state employment agencies. The core of the dataset is constructed by linking unemployment insurance records, which contain information on the job tenure and payment of employees, with firm-level data from the Quarterly Census of Employment and Wages. Details of the exact procedure are described in the academic papers about the creation of the datasets (Abowd et al. 2004, 2009a).

3.1 Coverage

The LEHD data includes all jobs covered by unemployment insurance and select federal government jobs. It excludes, most notably, self-employed individuals and federal employees working in defense-related agencies. While the exact proportion of private jobs covered by unemployment insurance is not known, the BLS estimated in 1997 that roughly 95% of private sector jobs are covered by unemployment insurance and thus included in the LEHD data (US Bureau of Labor Statistics 1997). The proportion covered by unemployment insurance may have declined with the rise of contract and contingent labor (Kalleberg 2000). Details on the exact jobs that are not covered by state unemployment are provided in (Stevens 2007). Details on the coverage of federal jobs are provided

at http://lehd.ces.census.gov/doc/help/onthemap/FederalEmploymentInOnTheMap.pdf and http://lehd.ces.census.gov/doc/help/onthemap/LODESDataNote-FedEmp2015.pdf.

3.2 Anonymization

Procedures are used to protect the anonymity of people covered by the LEHD files. In the QWI, LODES WAC, and J2J files, a multiplicative "fuzz factor" is generated for each employer and each establishment. This factor distorts the true estimates by a minimum of c and a maximum of d percent, where c and d are kept confidential (Abowd et al. 2009a). The fuzz factor assigned to each establishment is permanent, with the same factor used across different years and iterations of the data (Abowd et al. 2012). Further information on the details of the multiplicative noise generation process as applied to each of the datasets are available in the technical papers describing each of the datasets (Abowd et al. 2012, Abowd, McKinney 2016, Abowd et al. 2009a, Hyatt et al. 2017).

A more extensive anonymization process is used for the LODES RAC and OD files. Here, a full set of "synthetic data" is produced that attempts to preserve the statistical properties of the true data without actually being based on that data. Details of the construction of this synthetic data are described in (Abowd et al. 2012). Computer scientists have attempted to de-anonymize this data, and have described the limitations of the anonymization procedures used (Golle, Partridge 2009, Machanavajjhala et al. 2008).

3.3 Caveats

There are certain limitations to the LEHD datasets. First, as described above, the LEHD datasets do not fully cover the entire workforce, specifically lacking employees in defense-related industries.

Second, the geocoding of jobs to blocks in the LODES data is imperfect. Attempts are made to assign jobs at the establishment level, such that the jobs at a particular branch of a multi-branch company are located in the block containing that branch. However, this is not always possible. This difficulty is most prominent in the case of local government agencies like school districts and public transportation agencies. The jobs at such agencies are often assigned to the block containing their headquarters. This results in some blocks in the dataset having unreasonably high employment levels. The block containing the Brooklyn Municipal Building and the New York City Board of Education, for example, is listed as having 173,449 jobs in 2014. Studies of employment counts or density, especially those that use metrics sensitive to outliers, should take care to either drop or top-code these observations. Additionally, in certain types of jobs, such as home health aides, construction workers, and bus drivers, much of the work is performed at a location physically separate from the office, so even if employees are correctly assigned to the block containing their offices, this may not accurately reflect where the work is being done.

Third, the LODES data are not directly comparable across years, particularly at high geographic resolution. Improvements are continuously being made to the geocoding and job assignment algorithms, and these improvements are not typically applied backwards to previous data releases. The continuous improvement process means that the assignment process used in one year is not usually the same as the process used in the next year. In some cases these changes result in the movement of large numbers of jobs among neighboring blocks from year to year. Because of the continuously evolving algorithms, using the LODES data for longitudinal analysis at high spatial resolutions is not advised.

4 Examples of LEHD data use

The LEHD data products have begun to be used in a number of applications. The first applications were in labor economics without a geographic or spatial focus. For instance, Abowd et al. used the LEHD infrastructure file to examine how the relationship between technology and demand for skills varied within and across firms (Abowd et al. 2007). A number of studies have examined job to job mobility and flows into and out of employment (Abowd et al. 2009b, Abowd, Vilhuber 2011, Fallick et al. 2012, Haltiwanger et al. 2018). Some studies have examined earnings inequality (Abowd et al. 2018). There has also

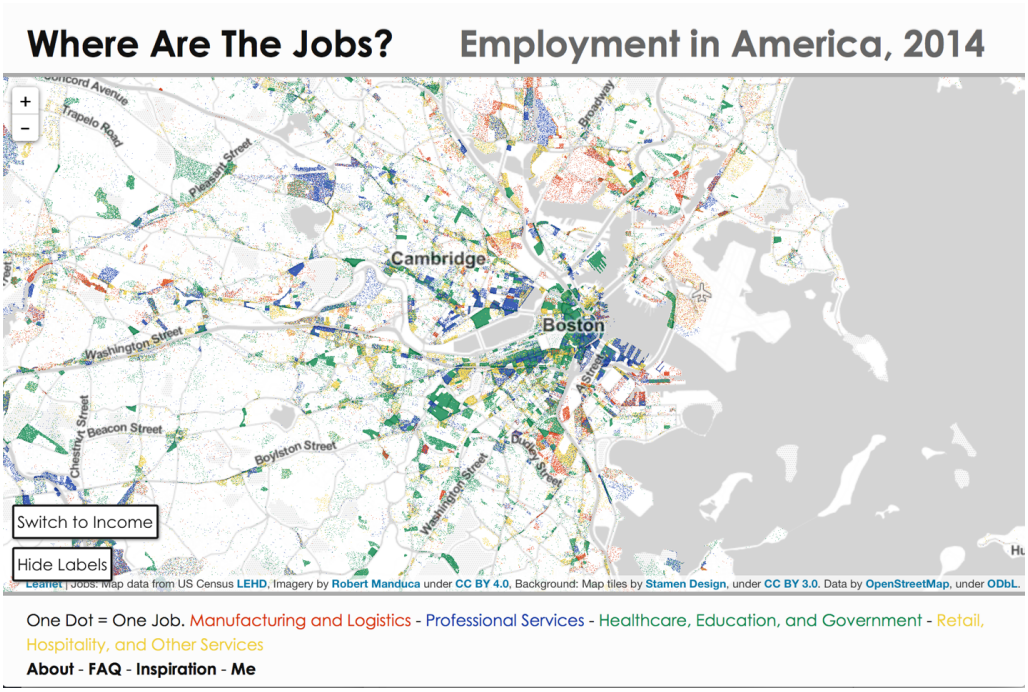

Figure 1: Screenshot from "Where Are the Jobs?", a web visualization of the LODES Workplace Area Characteristics data (http://www.robertmanduca.com/projects/jobs.html)

been increasing use of the LEHD data to study entrepreneurship (Goetz et al. 2015). Quarterly Workforce Indicators data have been used to study the effects of changes in the minimum wage (Dube et al. 2016).

Within regional science, the LODES origin-destination data have been used extensively to study commuting patterns and mobility (Horner, Schleith 2012, Horner et al. 2015, Kim 2014, Kneebone, Holmes 2015, Levinson, El-Geneidy 2009, Owen, Levinson 2015, Schleith, Horner 2014, Schleith et al. 2016). One aspect of this research includes studying vulnerability to emergencies (Kermanshah, Derrible 2016). The Workplace Area Characteristics files have been used to describe neighborhood character (Folch et al. 2017) and to identify business districts (Manduca 2018). The LODES data have also been used for web visualization, including the "Where Are The Jobs?" web maps (Manduca 2015).

The restricted use LEHD infrastructure files have been used to study spatial mismatch (Andersson et al. 2018), migration (Hyatt et al. 2016), and agglomeration (Freedman 2008). There is still a great deal of opportunity remaining for research using both the public and confidential data files.

5 Conclusion

The LEHD data products represent an impressive example of government agencies working across jurisdictions to create a truly new and rich source of information about regional economies. Although both the confidential and public versions of the data are being used more and more, there remains a huge amount of opportunity for future research. Regional scientists in the United States, Europe, and beyond should take advantage of this detailed and rich source of data.

References

Abowd J, Gittings RK, McKinney K, Stephens B, Vilhuber L, Woodcock S (2012) Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time series. Presented at the FCSM. http://digitalcommons.ilr.cornell.edu/ldi/5

Abowd J, Haltiwanger J, Lane J (2004) Integrated longitudinal employer-employee data for the United States. *American Economic Review* 94[2]: 224–229. CrossRef.

Abowd J, Haltiwanger J, Lane J (2009b) Wage structure and labor mobility in the United States. In: Lazear EP, Shaw KL (eds), *The Structure of Wages: An International Comparison*. University of Chicago Press, Chicago IL, 81–100. CrossRef.

Abowd J, Haltiwanger J, Lane J, McKinney KL, Sandusky K (2007) Technology and the demand for skill: An analysis of within and between firm differences. National Bureau of Economic Research, Cambridge. CrossRef.

Abowd J, McKinney K (2016) Noise infusion as a confidentiality protection measure for graph-based statistics. *Statistical Journal of the IAOS* 32[1]: 127–135. CrossRef.

Abowd J, McKinney KL, Zhao NL (2018) Earnings inequality and mobility trends in the United States: Nationally representative estimates from longitudinally linked employer-employee data. *Journal of Labor Economics* 36[S1]: S183–S300. CrossRef.

Abowd J, Stephens BE, Vilhuber L, Andersson F, McKinney KL, Roemer M, Woodcock S (2009a) The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In: T. Dunne, J. B. Jensen MJR (ed), *Producer Dynamics: New Evidence from Micro Data*. University of Chicago Press, Chicago, IL, 149–230. CrossRef.

Abowd J, Vilhuber L (2011) National estimates of gross employment and job flows from the quarterly workforce indicators with demographic and industry detail. *Journal of Econometrics* 161[1]: 82–99. CrossRef.

Andersson F, Haltiwanger JC, Kutzbach MJ, Pollakowski HO, Weinberg DH (2018) Job displacement and the duration of joblessness: The role of spatial mismatch. *Review of Economics and Statistics* 100[2]: 203–218. CrossRef.

Dube A, Lester TW, Reich M (2016) Minimum wage shocks, employment flows, and labor market frictions. *Journal of Labor Economics* 34[3]: 663–704. CrossRef.

Fallick B, Haltiwanger J, McEntarfer E (2012) Job-to-job flows and the consequences of job separations. Federal Reserve Board, Washington DC. https://www.federalreserve.gov/Pubs/feds/2012/201273/201273pap.pdf

Folch DC, Spielman SE, Manduca R (2017) Fast food data: Where user-generated content works and where it does not. *Geographical Analysis* 50[2]: 125–140. CrossRef.

Freedman ML (2008) Job hopping, earnings dynamics, and industrial agglomeration in the software publishing industry. *Journal of Urban Economics* 64[3]: 590–600. CrossRef.

Goetz C, Hyatt H, McEntarfer E, Sandusky K (2015) The promise and potential of linked employer-employee data for entrepreneurship research. National Bureau of Economic Research, Cambridge. CrossRef.

Golle P, Partridge K (2009) On the anonymity of home/work location pairs. In: Tokuda H, Beigl M, Friday A, Brush A, Tobe Y (eds), *Pervasive Computing. Pervasive 2009*, Volume 5538 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg. CrossRef.

Graham MR, Kutzbach MJ, McKenzie B (2014) Design comparison of LODES and ACS commuting data products. Working Paper No. 14-38, Center for Economic Studies, U.S. Census Bureau, Washington DC. https://ideas.repec.org/p/cen/wpaper/14-38.html

Green AS, Kutzbach MJ, Vilhuber L (2017) Two perspectives on commuting: A comparison of home to work flows across job-linked survey and administrative files. Working Paper No. 17-34, Center for Economic Studies, U.S. Census Bureau, Washington DC. https://ideas.repec.org/p/cen/wpaper/17-34.html

Haltiwanger J, Hyatt H, McEntarfer E (2018) Who moves up the job ladder? *Journal of Labor Economics* 36[S1]: S301–S336. CrossRef.

Horner MW, Schleith D (2012) Analyzing temporal changes in land-use-transportation relationships: A LEHD-based approach. *Applied Geography* 35[1-2]: 491–498. CrossRef.

Horner MW, Schleith DK, Widener MJ (2015) An analysis of the commuting and jobs-housing patterns of older adult workers. *The Professional Geographer* 67[4]: 575–585. CrossRef.

Hyatt H, McEntarfer E, Ueda K, Zhang A (2016) Interstate migration and employer-to-employer transitions in the US: New evidence from administrative records data. Working paper. https://papers.ssrn.com/sol3/papers.cfm?.abstract_id=2859095

Hyatt H, McKinney K, Vilhuber L, Hahn JK, McEntarfer E, Tibbets S, Walton D, Janicki H (2017) Job-to-job flows: New statistics on worker reallocation and job turnover. US Census Bureau, Washington DC. https://lehd.ces.census.gov/doc/-jobtojob_documentation_long.pdf

Kalleberg AL (2000) Nonstandard employment relations: Part-time, temporary and contract work. *Annual Review of Sociology* 26[1]: 341–365. CrossRef.

Kermanshah A, Derrible S (2016) A geographical and multi-criteria vulnerability assessment of transportation networks against extreme earthquakes. *Reliability Engineering & System Safety* 153: 39–49. CrossRef.

Kim JH (2014) Residential and job mobility: Interregional variation and their interplay in US metropolitan areas. *Urban Studies* 51[13]: 2863–2879. CrossRef.

Kneebone E, Holmes N (2015) The growing distance between people and jobs in metropolitan America. Brookings Institution, Washington DC. https://www.brookings.edu/-research/the-growing-distance-between-people-and-jobs-in-metropolitan-america/

Levinson D, El-Geneidy A (2009) The minimum circuity frontier and the journey to work. *Regional Science and Urban Economics* 39[6]: 732–738. CrossRef.

Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L (2008) Privacy: Theory meets practice on the map. IEEE 24th international conference on data engineering, Cancun, Mexico. CrossRef.

Manduca R (2015) Where are the jobs? http://www.robertmanduca.com/projects/jobs.-html

Manduca R (2018) The spatial structure of US metropolitan employment: New insights from LODES data. SocArXiv. September 13. CrossRef.

Owen A, Levinson DM (2015) Modeling the commute mode share of transit using continuous accessibility to jobs. *Transportation Research Part A: Policy and Practice* 74: 110–122. CrossRef.

Schleith D, Horner M (2014) Commuting, job clusters, and travel burdens: Analysis of spatially and socioeconomically disaggregated longitudinal employer-household dynamics data. *Transportation Research Record: Journal of the Transportation Research Board* 2452: 19–27. CrossRef.

Schleith D, Widener M, Kim C (2016) An examination of the jobs-housing balance of different categories of workers across 26 metropolitan regions. *Journal of Transport Geography* 57: 145–160. CrossRef.

Stevens DW (2007) Employment that is not covered by state unemployment insurance laws. US Census Bureau, Suitland MD. https://www2.census.gov/ces/tp/tp-2007-04.pdf

US Bureau of Labor Statistics (1997) Employment and wages covered by unemployment insurance, ch. 5. BLS Handbook of Methods (pp. 42–47). US Department of Labor, Washington DC. http://www.bls.gov/opub/hom/homch5_b.htm

US Census Bureau (2017) Job-to-job flows: New statistics on worker flows across jobs. Washington DC. https://lehd.ces.census.gov/doc/J2J_quickstart_guide.pdf