DESIGN COMPARISON OF LODES AND ACS COMMUTING DATA PRODUCTS

by

Matthew R. Graham
U.S. Census Bureau

Mark J. Kutzbach
U.S. Census Bureau

Brian McKenzie
U.S. Census Bureau

CES 14-38 September, 2014

Abstract

The Census Bureau produces two complementary data products, the American Community Survey (ACS) commuting and workplace data and the Longitudinal Employer-Household Dynamics (LEHD) Origin- Destination Employment Statistics (LODES), which can be used to answer questions about spatial, economic, and demographic questions relating to workplaces and home-to-work flows. The products are complementary in the sense that they measure similar activities but each has important unique characteristics that provide information that the other measure cannot. As a result of questions from data users, the Census Bureau has created this document to highlight the major design differences between these two data products. This report guides users on the relative advantages of each data product for various analyses and helps explain differences that may arise when using the products.2,3 As an overview, these two data products are sourced from different inputs, cover different populations and time periods, are subject to different sets of edits and imputations, are released under different confidentiality protection mechanisms, and are tabulated at different geographic and characteristic levels. As a general rule, the two data products should not be expected to match exactly for arbitrary queries and may differ substantially for some queries. Within this document, we compare the two data products by the design elements that were deemed most likely to contribute to differences in tabulated data. These elements are: Collection, Coverage, Geographic and Longitudinal Scope, Job Definition and Reference Period, Job and Worker Characteristics, Location Definitions (Workplace and Residence), Completeness of Geographic Information and Edits/Imputations, Geographic Tabulation Levels, Control Totals, Confidentiality Protection and Suppression, and Related Public-Use Data Products. An in-depth data analysis—in aggregate or with the microdata—between the two data products will be the subject of a future technical report. The Census Bureau has begun a pilot project to integrate ACS microdata with LEHD administrative data to develop an enhanced frame of employment status, place of work, and commuting. The Census Bureau will publish quality metrics for person match rates, residence and workplace match rates, and commute distance comparisons.

Design Comparison of LODES and ACS Commuting Data Products

By Matthew R. Graham (LEHD), Mark J. Kutzbach (LEHD), Brian McKenzie (SEHSD)[1]
Issued October 2014

Introduction/Overview

The U.S. Census Bureau produces two complementary data products, the American Community Survey (ACS) commuting and workplace data and the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES), which can be used to answer questions about spatial, economic, and demographic questions relating to workplaces and home-to-work flows. The products are complementary in the sense that they measure similar activities but each has important unique characteristics that provide information that the other measure cannot. As a result of questions from data users, the Census Bureau has created this document to highlight the major design differences between these two data products. This report guides users on the relative advantages of each data product for various analyses and helps explain differences that may arise when using the products.[2, 3]

As an overview, these two data products are sourced from different inputs, cover different populations and time periods, are subject to different sets of edits and imputations, are released under different confidentiality protection mechanisms, and are tabulated at different geographic and characteristic levels. As a general rule, the two data products should not be expected to match exactly for arbitrary queries and may differ substantially for some queries.

Within this document, we compare the two data products by the design elements that were deemed most likely to contribute to differences in tabulated data. These elements are: Collection, Coverage, Geographic and Longitudinal Scope, Job Definition and Reference Period, Job and Worker Characteristics, Location Definitions (Workplace and Residence), Completeness of Geographic Information and Edits/Imputations, Geographic Tabulation Levels, Control Totals, Confidentiality Protection and Suppression, and Related Public-Use Data Products.

An in-depth data analysis—in aggregate or with the microdata—between the two data products will be the subject of a future technical report. The Census Bureau has begun a pilot project to integrate ACS microdata with LEHD administrative data to develop an enhanced frame of employment status, place of work, and commuting. The Census Bureau will publish quality metrics for person match rates, residence and workplace match rates, and commute distance comparisons.

Document Outline

The remainder of this document is organized as follows:

- Data Product Overview
- Detailed Design Differences
- Future Plans
- Key References
- Appendix: Design Difference Summary Table

[1] LEHD refers to Longitudinal Employer-Household Dynamics and is a program area of the Center for Economic Studies. SEHSD refers to Social, Economic, and Housing Statistics Division.

[2] This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. This paper can be found at <ftp://ftp2.census.gov/ces/wp/2014/CES-WP-14-38.pdf>.

[3] The authors would like to thank the following people for their assistance in preparing this report: Lucia Foster, Henry Hyatt, William Koerber, James Spletzer, Stephen Tibbets, Douglas Walton, and Robert Sienkiewicz.

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

United States™ Census Bureau

Data Product Overview

Commuting statistics (or Journey-to-Work) are tabulated from the ACS,[4] an ongoing survey of the Population of the United States conducted by the U.S. Census Bureau. The ACS questions related to travel focus solely on commuting and do not ask about nonwork travel. Respondents answer questions about where they work, what time they leave home for work, the means of transportation used to get there, the number of workers riding in the car, truck, or van, and how long it takes to travel to work. Along with several other questions on housing and person characteristics, these travel-related questions previously appeared on the decennial census long form, but were moved to the ACS after the 2000 Census. The decennial census now produces a count of the nation's population and a snapshot of its most basic demographic and housing characteristics.

The LODES are produced by the LEHD program[5] at the U.S. Census Bureau. The LODES data are an extract of the LEHD data infrastructure, which is composed of administrative records, census and survey data focused on the labor market, worker, and firm statistics. State unemployment insurance reporting and account information and federal worker earnings records provide information on employment location for covered jobs and residential information for workers, which form the basis of the LODES data product. However, these data are not available in all states for all years of the series.

Detailed Design Differences

The following areas of comparison were chosen for inclusion in this report because they were deemed most likely to assist data users in understanding the differences between ACS commuting data and LODES. The areas covered in this section are:

- Collection
- Coverage
- Geographic and Longitudinal Scope
- Job Definition and Reference Period
- Job and Worker Characteristics
- Location Definitions (Workplace and Residence)
- Completeness of Geographic Information and Edits/Imputations
- Geographic Tabulation Levels
- Control Totals
- Confidentiality Protection and Suppression
- Related Public-Use Data Products

Collection

The ACS is a nationwide survey designed to provide communities with reliable and timely demographic, social, economic, and housing data for congressional districts, counties, places, and other localities every year. It has an annual sample size of about 3.5 million addresses across the United States and Puerto Rico and includes both housing units and group quarters (e.g., military barracks, nursing homes, and prisons). The ACS is conducted in every county throughout the nation and every municipio in Puerto Rico, where it is called the Puerto Rico Community Survey (PRCS). Beginning in 2006, ACS data for 2005 were released for geographic areas with populations of 65,000 and greater.[6]

LODES data are generated from a variety of administrative and survey data. The core jobs data come from states that are part of the Local Employment Dynamics (LED) Partnership. The LED Partnership is a voluntary federal-state partnership that was started in 1999. Its main purpose is to merge data from workers

[4] ACS program Web site <www.census.gov/acs/www/>.
[5] LEHD program Web site <http://lehd.ces.census.gov/>.
[6] For information on the ACS sample design and other topics see: U.S. Census Bureau, Design and Methodology, American Community Survey, U.S. Government Printing Office, Washington, DC, 2009, <www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf>.

with data from employers to produce a collection of enhanced labor market statistics known collectively as Quarterly Workforce Indicators (QWI), subject to strict protection of the identity and confidentiality of the individual respondents.[7] States joined at different times and the extent of time series availability varies by state. By 2013, all states, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands had joined LED. These records are supplemented by earnings records of federal workers from the Office of Personnel Management (OPM).

Coverage

The ACS estimates are derived from a sample of the population in the United States and Puerto Rico. The ACS and the PRCS each consist of two separate samples: housing unit (HU) addresses and persons in group quarters (GQ) facilities. The sampling frames from which these samples are drawn are derived from the Census Bureau's Master Address File (MAF). The MAF is the Census Bureau's official inventory of known living quarters and selected nonresidential units in the United States and Puerto Rico. Independent HU address samples are selected for each of the 3,142 counties and county equivalents in the United States, including the District of Columbia, for the ACS. Similarly, for the PRCS, address samples are selected for each of the 78 municipalities in Puerto Rico.[8] Questions about transportation pertain only to work travel. ACS estimates related to commuting are derived from a sample of workers 16 years and over who worked during the ACS reference week, the calendar week preceding the date respondents completed their questionnaire.

LODES provides counts of unemployment insurance covered wage and salary jobs, as reported by state labor market information offices and by OPM (for 2010 onwards). The state data, covering employers in the private sector and state and local government, account for approximately 95 percent of wage and salary jobs.[9] Partner states submit two quarterly files from their administrative data systems. Wage record files list the quarterly earnings for workers at each job they hold in a pay period falling within the quarter. These files have no information on hours or weeks worked. Quarterly Census of Employment and Wages (QCEW) files, formerly known as the ES-202 program, list the establishments at employers active in a quarter, and report location, industry, ownership type, and workforce size. LEHD does not receive or process Multiple Worksite Reports (listing the establishments associated with an employer), which states have already incorporated into QCEW files. The OPM data cover most federal government civilian employment.[10] Within this universe of jobs, the LODES data have almost full coverage, although some in-universe firms may not report during a particular quarter. Examples of job types beyond the scope of LEHD earnings records are: the military and other security-related federal agencies, postal workers, some employees at nonprofits and religious institutions, informal workers, and the self-employed.

Geographic and Longitudinal Scope

The Census Bureau has produced commuting tables from survey data since the 1960 Census. The 2000 Census was the last decennial census to include a "long form" sample and the last to ask questions on commuting. Information about commuting is now collected solely through the ACS. Data collection for the ACS was first conducted in 1996 for four test sites. The number of test sites was increased during the development stage of the survey. The demonstration stage started with the 2000 ACS where the survey sample was expanded to 1,239 counties across the United States. Full implementation began with the 2005 ACS when the survey was expanded to all counties across the United States and Puerto Rico. Group quarters collection started with the 2006 ACS.

[7] For more on the processing of jobs data and production of the QWI, see J.M. Abowd et al., "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators," LEHD Technical Paper No. TP-2006-01, 2005, <http://lehd.ces.census.gov/doc/technical_paper/tp-2006-01.pdf>.

[8] For Coverage Rates Data, see <www.census.gov/acs/www/methodology/coverage_rates_data/>.

[9] For more on coverage, which varies across states, see D. W. Stevens, "Employment that is not covered by state unemployment insurance Laws," LEHD Technical Paper No. TP-2007-04, 2007, available at <http://econpapers.repec.org/paper/centpaper/2007-04.htm>.

[10] Federal employment first appears in 2010 LODES data. For more information on coverage within the civilian federal workforce, see Data Definitions: <www.fedscope.opm.gov/datadefn/aboutehri_sdm.asp#cpdf3> and <www.fedscope.opm.gov/datadefn/index.asp#location>.

LODES are published as an annual cross-section from 2002 onwards, with each job having a workplace and residence dimension. States provide data for employers located within state, so only states contributing data to LEHD in a given year will be included in the workplace dimension for that year. The release of 2011 jobs data (the most recent release at the time of writing) includes all states except Massachusetts, Puerto Rico, and the U.S. Virgin Islands. The federal administrative data providing residence locations are national and do not depend on a state joining the LED Partnership. LODES do not currently release residence data for Puerto Rico and the U.S. Virgin Islands. While the domain of workplace locations varies across years, all states have some—possibly incomplete—residence data from 2002 onwards.

The LEHD program produces each year of LODES independently, so there may be time series inconsistencies due to updates and methodological changes that can complicate longitudinal inferences. LODES processing always waits for at least two quarterly submissions from each state before processing data for a year, but additional updates may follow. (See "Job Definition and Reference Period" section for more detail on the reference period.) Updates might include additional earnings records or new information on existing records. Quarterly processing of the LEHD infrastructure files also updates imputations for missing values such as establishment assignments to workers. Subsequent LODES releases reflect the data at the time of extract and data already released for earlier years do not incorporate these updates.

One example of a change in processing is in the stratifying distribution used for synthesizing residence locations (see below for more discussion on confidentiality protection). From 2002 to 2008, the distribution of residences for a workplace was stratified based on a worker's age, earnings, and industry group. Beginning in 2009, the stratifying variables were broadened to include ownership, industry group, earnings, age, sex, race, and ethnicity.

Job Definition and Reference Period

For ACS commuting tables, workers must be aged 16 or older and answer in the affirmative to the questions, "LAST WEEK, did this person work for pay at a job (or business)?" or "LAST WEEK, did this person do ANY work for pay, even for as little as one hour?"[11] ACS respondents are asked to report information about their primary job (the one where they worked the most hours in the last week). If the respondent worked at more than one location, they are to indicate the job where they worked the most hours. No secondary jobs are reflected in ACS commuting estimates. For respondents to be included in workplace or commuting tabulations, they must also have a discernible response to the question, "At what location did this person work LAST WEEK?" The universe of workers is limited to those 16 years and over who worked during the ACS reference week, the calendar week preceding the date respondents completed their questionnaire. Workers who were on vacation or sick leave the entire reference week are not included. Workers living in noninstitutional group quarters are included. Those living in institutional group quarters (e.g., prisons, nursing homes) are not included.

The ACS reference week is the calendar week preceding the date respondents completed their questionnaire. Additional characteristics specific to the job held in the last week are: industry, occupation, and class of worker. Hours worked each week, weeks worked in the past 12 months, and earnings may refer to all jobs over the last 12 months. ACS 1-, 3-, and 5-year estimates are period estimates, which mean they represent the characteristics of the population and housing over a specific data collection period. Data are combined to produce 12 months, 36 months, or 60 months of data. Commuting data based on decennial censuses (2000 and prior) have a reference date of April 1 for each census year.

The LEHD infrastructure collects earnings records and QCEW data quarterly. For a job to be included in LODES, a person must have positive earnings from a firm in both the reference quarter (Quarter 2 of each year: April to June) and the previous quarter (Quarter 1: January to March). Jobs with earnings in adjacent quarters are inferred to have been held at the seam of the quarters, in this case, April 1. Because quarterly earnings reports often depend upon completion of a pay period during a quarter, LODES may underrepresent very short jobs held for just one pay period. The LODES job count corresponds to the QWI indicator "beginning-of-quarter" jobs, for the second quarter.

[11] ACS questionnaires can be accessed at <www.census.gov/acs/www/methodology/questionnaire_archive/>.

While all jobs in the LODES snapshot are presumed to be held on April 1, some job characteristics may have a different timeframe. Earnings for a job include all income from an employer during all pay periods falling within the quarter. LODES uses second quarter nominal earnings to classify jobs into low, middle, and high earning categories. LODES further classifies jobs as "primary/dominant" if the job earned the person the most earnings of all jobs held during the reference period. Other employer characteristics, including ownership, industry, and firm size and age are also based on the employer report in the second quarter. The timing of the residential location is not as precise, but use of tax filing data means that many residential locations have a reference date on or before April 15 of each year. Worker demographics are assumed to be time invariant and are based upon survey and administrative records. Educational attainment is not available for those in the youngest age category (29 or younger).

Job and Worker Characteristics

Standard ACS commuting tables include commuting characteristics such as travel time, time of departure from home to work, travel mode, and workplace location. These are crossed with several demographic and job-related variables such as age, sex, race, ethnicity, poverty status, industry occupation, class of worker and others. Detailed ACS commuting tables have the prefix "B08" and are available on American FactFinder (AFF). Tables with the prefix "B080" contain worker characteristics for residence geography; "B081" tables consist of means of transportation crossed by demographic, social, economic, and housing characteristics for residence geographies; "B082" tables are household tables; "B083" tables are basic commuting tables for residence geographies; "B084" tables are worker characteristics for workplace geography; "B085" tables consists of means of transportation crossed by demographic, social, economic, and housing characteristics for workplace geographies; and "B086" tables are basic commuting tables for workplace geographies.

LODES tabulates jobs by workplace and residence location, as well as by a set of job, employer, and worker characteristics. All tabulations are available by job dominance (2 categories) and employer ownership (3 categories). See the "Job Definition" section above for further explanation of job dominance. Ownership includes federal government, state and local government, and the private sector, although the federal job type only appears in raw data files and only for 2010 and later. Other characteristics are industry (3 industry groups, 20 NAICS sectors), firm size (5 categories), firm age (5 categories), average monthly earnings (3 categories), age (3 categories), sex (2 categories), race (6 categories), ethnicity (2 categories), and educational attainment (4 categories, and not applicable for those aged 29 and under).[12, 13]

Not all characteristics are available throughout the time series. Sex, race, ethnicity, and educational attainment are only available beginning in 2009. Firm size and age are only available beginning in 2011.[14] Federal data from OPM begin in 2010. Earnings categories are for nominal values that are constant over the span of the series. The three age categories are not aligned with the eight age categories in QWI.

Location Definitions (Workplace and Residence)

The ACS questionnaire specifically asks for the address where a person worked last week. A job within the commuting data is defined by the question from the ACS questionnaire, "At what location did this person work LAST WEEK?" Respondents are encouraged to report the physical location of their workplace establishment rather than an administrative address. Enumerators instruct respondents to report last week's location, even if it was not their regular office or administrative office, and even if it was just a business trip or conference. As a result, some reported workplace locations are not typical and do not correspond logically to residence locations (i.e., where the survey was sent) based on reported travel times within the microdata. For example, a worker may report leaving home in New York and report working in Chicago, with a walking-based travel time of 10 minutes. In this case, the respondent may be attending a conference in a different city.

[12] NAICS is the North American Industrial Classification System.

[13] Additional detail on the variables provided in LODES can be found at "OnTheMap: Data Overview (LODES Version 7)" (2013), <http://lehd.ces.census.gov/doc/help/onthemap/OnTheMapDataOverview.pdf>.

[14] For more on firm age and size, see "Quarterly Workforce Indicators: New Jobs Data by Firm Age and Firm Size," 2012, <http://lehd.ces.census.gov/doc/FirmAgeAndSizeOnePager.pdf>.

LODES documentation often uses the terms "origin-destination flow" and "commute" to refer to the residence and workplace tied to a job. However, these flows may not always represent a trip pattern. As is discussed below, both the workplace and residence may differ from a typical travel survey concept. The LEHD data infrastructure and LODES in particular have no information about transport mode (if any), and a worker may telework from home regularly or work only infrequently.

For LODES, a place of work is defined by the physical or mailing address reported by employers in the QCEW (formerly ES-202) or Multiple Worksite Reports.[15] An address from administrative data may or may not be the actual location that a worker reports to most often. The distinction of worksite and administrative address may be especially significant in some industries such as construction, where work is often carried out at temporary locations. In some cases, employers do not provide a multiple worksite report when it would be appropriate to do so. Nonreporting of multiple worksites is especially common with state and local governments and school districts. In such a case, LEHD infrastructure files assign all workers for that employer (within the state) to the main address provided. Bureau of Labor Statistics (BLS) data show a national noncompliance rate of 5.61 percent of multiunit employers responsible for about 4.45 percent of multiunit employment.[16]

The residence location for workers in LODES is derived from federal administrative records. For the years 1999 to 2010, the Census Bureau produced a file known as the Composite Person Record (CPR) that listed a single residence location, geocoded to a census block, for a set of workers in each year. The Center for Economic Studies and LEHD now produce a similar file from the same sources. LEHD infrastructure files use the CPR for imputation models requiring place of residence, and LODES uses it for the likelihood distribution in the residence synthesizer. If a worker moves during the year, or lives and works in a different area than the residence in the federal data, the associated "commute" may appear to be a greater (or shorter) distance than the worker actually travels (if at all) from home on a regular basis.

Completeness of Geographic Information and Edits/Imputations

The ACS geocoding operation first attempts to code workplace location to the place level, then to the block level. For some worker records, the Census Bureau is unable to code the workplace address to the block level. This may occur for a variety of reasons, including incomplete workplace address information provided from the respondent, or inability of the geographic coding system to find a match between the employer name and street address responses. When this is not possible due to insufficient address information, a respondent's workplace location is imputed to the place level. In 2012, about 92 percent of records were successfully coded to the place level.[17] For records that cannot be coded down to the place level, workplace location is imputed using a series of characteristics such as industry, travel mode, and travel time from records with known workplace addresses. The Census Bureau attempts to code down to the block level, but does not display estimates at the workplace geographies smaller than the place level for standard ACS tables.

Workplace geography is coded to the latest census tabulation geography (i.e., currently 2010 Census tabulation geography). During the edit process, it is updated to the same vintage geography for current residence. When the multiyear ACS files are created, the workplace geography for prior years is again updated to the vintage of the most recent year. The geography for current residence usually uses legal boundaries as of January 1 of the last year of the estimate period.

For LODES, imputations begin in the LEHD infrastructure files, where missing or incomplete employment, geography, industry, and demographic information are imputed using models developed from jobs with complete information. In most cases, LODES data processing uses LEHD infrastructure-file imputation models to complete job and worker characteristics. One exception is for workplace and residence geography variables, for which LODES requires greater precision. Both the LEHD infrastructure files (used for QWI)

[15] The BLS provides links to all states' Multiple Worksite Reports at <www.bls.gov/cew/cewmwr00.htm>.

[16] See Table 2.1 of Bruce D. Spear, "Improving Employment Data for Transportation Planning," NCHRP 08-36, Task 098, 2011, <www.camsys.com/pubs/NCHRP08-36-98.pdf>.

[17] See AFF Table B99081, Imputation of Place of Work.

and LODES use multiple imputation, with all missing values being completed using ten independent draws from a predictive distribution and reweighted by one-tenth (for tabulation). Multiple-imputation helps to reflect the uncertainty of the underlying data.

Approximately 40 percent of LEHD jobs are at employers reporting multiple establishments (those filing a Multiple Worksite Report), except for Minnesota employers who do not assign workers to an establishment. The LEHD program uses an imputation model with parameters based on the Minnesota data to draw establishments for workers at multiunit employers. An establishment is more likely to be assigned to a worker when it is large and close to that worker's residential location (based on great-circle distance between address coordinates).

The LEHD program uses a custom geocoder to assign incoming business addresses to census geography and latitude/longitude coordinates. Approximately 97 percent of jobs are at an establishment where the address has subcounty precision. Another 2 percent have county level geographic information and the remaining jobs have no valid workplace address (although the workplace is assumed to be located in the state that provided the earnings records). LODES retains the geographic quality indicators from LEHD infrastructure files, but carries out its own—more geographically precise—imputation. For each job with an imprecise workplace location, LODES imputes missing geography by drawing from the distribution of workplaces for coresidents (residential neighbors) who are employed in the same NAICS industry group.

Approximately 96 percent of workers have a residence with at least county precision. Before creating the likelihood distribution for the synthetic data model, all residence addresses are completed to census tract level precision, using a distribution based on the observed residences of coworkers who fall in the same quarterly earnings category.

Geographic Tabulation Levels

ACS 1-, 3-, and 5-year estimates are period estimates, which means they represent the characteristics of the population and housing over a specific data collection period. Data are combined to produce 12 months, 36 months, or 60 months of data. These are called 1-year, 3-year, and 5-year ACS datasets. The 1-year data are released for areas of 65,000 populations or greater, 3-year estimates are released for areas of 20,000 populations or greater, and 5-year estimates are released for all areas. The 5-year ACS standard commuting tables are released down to the block group level for residence-level geography (where people live) and to the place level for workplace geography (where people work). When the multiyear ACS files are created, the workplace and residence geography for prior years is updated to the vintage of the most recent year.

Standard ACS tables present information about where people live and where they work, but information about the residence/workplace relationship is not provided as an origin-destination combination. Although not included among standard ACS data products, commuting flow estimates are provided as part of the Office of Management and Budget's (OMB's) effort to delineate the nation's metropolitan (metro), micropolitan (micro), and related statistical areas using ACS data. These are county-level and minor civil division (MCD)-level commuting flows, most recently produced from the 2006–2010 ACS 5-year estimates.[18]

LODES are produced and released at the census block level, with all tabulations consisting of paired, origin-destination flows that can also be aggregated to the residence and workplace margins. LODES documentation refers to these flows as Origin-Destination Tables, and the margins as Workplace Area Characteristics and Residence Area Characteristics Tables.[19] All higher-level aggregations produced by the OnTheMap application are sums of whole census blocks. The mapping of census blocks to higher-level geography is based on the internal point of a block falling within the boundary of the higher geography, except when the block's inclusion in a statistical area is mandated by law.[20] Since the release of 2011 data

[18] For more information on county-level commuting flows, see <www.census.gov/hhes/commuting/data/commuting.html>.

[19] For a complete description of the LODES data structure, see "OnTheMap: Data Overview (LODES Version 7)" (2013), <http://lehd.ces.census.gov/doc/help/onthemap/OnTheMapDataOverview.pdf>.

[20] Typically this happens only in relation to political redistricting. For specific examples from 2012 state legislative redistricting as well as 113th Congress redistricting, see the "Split Block Reports" section of <www.census.gov/rdo/data/113th_congressional_and_2012_state_legislative_district_plans.html>. For more general information about redistricting, see <www.census.gov/rdo/>.

(LODES Version 7), all tabulations of census geography in all years (2002 to present) are based on 2010 Census tabulation geography. Because earlier releases of LODES were produced for 2000 Census tabulation geography, older years had to be "cross-walked" to the present definitions.[21]

Control Totals

The ACS uses independently derived annual population estimates as population controls in its weighting methodology.[22] The 2010 ACS estimates and later years use population estimates that reflect the results of the 2010 Census. Population controls are applied down to the county level for basic demographic characteristics such as age, sex, and race. Population totals without characteristics are applied for some subcounty geographic summary levels, specifically places and minor civil divisions.

At the state level, the source data for LODES are reweighted to state totals released by the BLS as the QCEW. LODES is not necessarily consistent with QWI beginning of quarter job counts at any level. The LODES extraction from the LEHD infrastructure files and the definition of jobs have not been completely consistent with QWI tabulations across years. Differences in geographic imputations and small cell protections lead to other differences. The LODES and QWI age categories are not synchronized.

Confidentiality Protection and Suppression

The Census Bureau protects confidentiality by applying disclosure techniques and limiting the geographic areas of release for ACS tables.[23] Data swapping is the main procedure used for ACS tabulations. In each case, a small percentage of household records are swapped. Pairs of households in different geographic regions are swapped. The selection process for deciding which households should be swapped is highly targeted to affect the records with the most disclosure risk. Pairs of households that are swapped match on a minimal set of demographic variables. All data products are created from the swapped data files.[24] There are also limits on the number of variables to be crossed in a single table. Tables for workplace geography crossed by characteristics are suppressed if the universe contains less than 50 unweighted cases. Another form of data suppression occurs as a result of data quality concerns. There are several standards related to the number of cases included in a cell and the relationship between the estimate and margin of error.[25]

For Public Use Microdata Sample (PUMS), the geographic codes available on the PUMS files are limited to regions, divisions, states, and Public Use Microdata Area (PUMA), which has a population threshold of 100,000 or more. There are also Workplace Public Use Microdata Areas (POWPUMA). Each POWPUMA is equivalent to a PUMA or group of PUMAs. Additional disclosure techniques applied to the PUMS files are top-coding of variables, age perturbation, and collapsing of categorical variables.

LODES uses noise infusion and small cell imputation methods to protect workplace job counts and synthetic data methods to protect the residential location of jobs. The protection of workplace counts uses the same procedure as the QWI, namely, multiplying job counts by randomly generated "fuzz factors" specific to each employer and establishment. Although earnings records inherently are counts of whole jobs, multiple imputation, weighting factors,[26] and the fuzz factors lead to fractional job counts. LODES rounds

[21] For more information on this cross-walking methodology, see "OnTheMap: Updating Base Geography" (2012), located at <http://lehd.ces.census.gov/doc/help/onthemap/OnTheMap2010Geography.pdf>. Also useful are the TIGER block relationship files, which are located at <www.census.gov/geo/maps-data/data/relationship.html>.

[22] For information about population controls, see American Community Survey Research Note, "Change in Population Controls," September 22, 2011, <www.census.gov/acs/www/Downloads/comparing_acs_data/2010_Change_Population_Controls.pdf>.

[23] See Laura Zayatz, "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update," Statistical Research Division Research Report Series Statistics #2005-06, 2005, <www.census.gov/srd/papers/pdf/rrs2005-06.pdf>.

[24] See Chapter 13, "Preparation and Review of Data Product - Disclosure Avoidance section of U.S. Census Bureau," Design and Methodology, American Community Survey, U.S. Government Printing Office, Washington, DC, 2009, <www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology_ch13.pdf>.

[25] For a detailed explanation of data suppression rules relevant to ACS estimates, see U.S. Census Bureau, "American Community Survey: Data Suppression," 2013, <www.census.gov/acs/www/Downloads/data_documentation/data_suppression/ACSO_Data_Suppression.pdf>.

[26] Weighting factors account for differences in earnings record-reported job counts compared to QCEW-reported job counts from employers, and they also reweight statewide QWI job counts to match the totals released by BLS.

these job counts to the nearest integer at the workplace cell level, with a workplace tabulation cell defined by a census block and a combination of job characteristics.

While QWI suppresses job counts in some cells due to excessive distortion, LODES never suppresses jobs. For cells with only one or two jobs following the fuzzing and rounding described above, LODES imputes a one or two for that cell based on the overall distribution of ones and twos in the census tract. Because there is no suppression at this most granular tabulation, there is also no suppression in more aggregate tabulations.

LODES synthesizes the residential location of workers, using a method with probabilistic differential privacy protections.[27] For each job in a workplace cell, LODES draws from a Dirichlet multinomial posterior distribution of possible residential locations. The domain of possible residences for a workplace cell includes the entire United States. LODES coarsens the residence domain depending on the distance of a residence from a workplace.[28] At minimum, all residences are coarsened to the census tract level. More distant residences may be coarsened up to PUMA or Super-PUMA target areas.[29]

The probability of drawing any residence depends on both a likelihood, which is based on the number of observed job flows from each coarsened residence to the workplace, and a prior distribution, which is more evenly distributed. The likelihood distribution is stratified by job characteristics (ownership, industry, earnings, age, sex, race, and ethnicity) of the workplace census tract that includes the workplace destination block, giving each type of worker a distinct distribution of residences. The prior adds uncertainty, so that even commutes with few or no observed flows may appear to have a job.[30] Conversely, even when there are commuters from an origin in the likelihood, that residence may not be drawn and thus would not appear in LODES.

Following synthesis, LODES imputes an even more detailed residence for all jobs. Jobs with an origin synthesized to the Super-PUMA or PUMA level are first decoarsened to the census tract level. Then all jobs are decoarsened to census blocks. The distribution for decoarsening is constructed from the population distribution, by race and ethnicity, for the previous decennial census (2000 or 2010). Lastly, the jobs are aggregated to the tabulation level defined by geography and sets of characteristics and released as LODES.

Related Public-Use Data Products

Several commuting tables are included in standard Census Bureau products and are available for residence and workplace geographies on the AFF Web site.[31] Tables prior to full implementation of the ACS in data year 2005 are available on the Census Bureau's FTP site.[32] The Census Bureau produces county-to-county commuting flows as part of the routine update of metropolitan and micropolitan statistical area boundaries.[33] The Census Bureau also provides PUMS data from the ACS in order for users to create their own tabulations. PUMS data include location information on the residence and workplace of ACS respondents.[34] In addition, the Census Bureau releases ACS-based reports on various commuting topics. Many of these reports are associated with supplemental tables not found on AFF.[35]

[27] For a description of the theory behind the LODES synthetic data model, see Ashwin Machanavajjhala et al., "Privacy: Theory meets Practice on the Map," ICDE 2008: 277–286.

[28] LODES uses great-circle distance calculated between internal points of home and work geographies to determine coarsening. Internal points are not mathematical centroids, but are merely guaranteed to be inside the boundary they represent. The coarsening is tuned so that, in general, the domain of 90 percent of home to work flows for a given workplace will not have residence location coarsened above the census tract level.

[29] PUMAs are geographic areas for which the Census Bureau provides selected extracts of raw data from a small sample of census records that are screened to protect confidentiality. These extracts are referred to as PUMS files. PUMAs have a minimum population of 100,000 residents, and Super-PUMAs have a minimum population of 400,000 residents.

[30] For the first release of LODES for any state, the prior is based on the distribution of commute patterns in the year 2000 CTPP data. In subsequent years, the prior is based on the previous release of LODES for that state.

[31] See <http://factfinder2.census.gov/>.

[32] See <http://www2.census.gov/>.

[33] For more information on county-to-county commuting flows, see Metropolitan and Micropolitan Statistical Areas Main: <www.census.gov/population/metro/>, and Commuting (Journey to Work): <www.census.gov/hhes/commuting/data/commuting.html>.

[34] See Public Use Microdata Sample (PUMS): <www.census.gov/acs/www/data_documentation/pums_data/>.

[35] See Commuting (Journey to Work) Main: <www.census.gov/hhes/commuting/>.

ACS-based worker estimates of commuting flows have served as the basis for several special data tabulations. Such tabulations are not considered to be standard Census Bureau products and may differ from standard Census Bureau tables in summary levels included in the tabulations and their treatment for confidentiality protection. For example, the Census Transportation Planning Products (CTPP) is the most detailed tabulation of ACS commuting data. It is a collaborative effort among the U.S. Census Bureau and several transportation-related agencies to produce a set of tabulations designed for transportation planners. This tabulation includes several commuting flow tables at geographic summary levels smaller than the county.[36] Much of the ACS/LODES discussion presented here on data collection, coverage, and job and location definitions will also apply to the source data for CTPP, but additional questions of imputation and suppression may apply.[37] The Equal Employment Opportunity (EEO) data are another example of a special tabulation involving commuting flow data. EEO flows are produced to monitor the effects of antidiscrimination labor laws such as the Equal Employment Opportunity laws. In a typical application of this sort, the demographic characteristics of the commuter shed are considered relative to those of the workforces of local firms.[38]

The LEHD program produces two public-use data products from a common set of infrastructure files. The infrastructure files enable linking of earnings records to employer and worker information, and include imputation processes to complete missing or invalid data. The LODES are made available in the On-The-Map Web tool, which enables users to aggregate jobs data for standard and custom geographic queries as well as for different sets of job characteristics.[39] The QWI provide tabulations of jobs data aggregated to workplace county, metropolitan area, and Workforce Investment Act geographies as well as for different sets of job characteristics.[40] LODES provides data on an annual cycle (see "Job Definition and Reference Period" section for more detail) with greater geographic detail along with residence information, while the QWI provide quarterly data with more characteristic detail. LODES only tabulates the count of jobs in predefined categories while QWI provides other indicators including worker flows and average earnings. LEHD releases LODES on an annual basis, with no updates to previous releases, while the entire QWI time series updates each quarter, using the latest data and methodology.

Future Plans

The Census Bureau has begun a pilot project to integrate ACS microdata with LEHD administrative data to develop an enhanced frame of employment status, place of work, and commuting. A person-level integration will enable a comprehensive investigation of how microdata differ between the two datasets, resulting in better understanding of the relative strengths of each. The project will result in the publication of quality metrics for person match rates, residence and workplace match rates, and commute distance comparisons, but could develop further as new priorities are identified. The Census Bureau team, including staff involved in the production of both LODES and ACS data, will provide guidelines and recommendations for users. The linking methodology will facilitate ongoing comparisons as well as the potential for data quality enhancements drawing on the strengths of both datasets.

Suggested Citation

Graham, Matthew R., Mark J. Kutzbach, and Brian McKenzie, "Design Comparison Between LODES and ACS Commuting Data Products," Working Paper CES-WP-14-38, Center for Economic Studies, U.S. Census Bureau, 2014.

[36] For further discussion of ACS and CTPP, see National Cooperative Research Program, "A Guidebook Using American Community Survey Data for Transportation Planning," NCHRP Report 588, 2007, <http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_588.pdf>. CTPP Data Products are accessible at <www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/index.cfm>.

[37] For a comparison of LODES and the CTPP estimates, see Bruce D. Spear, "Improving Employment Data for Transportation Planning," NCHRP 08-36, Task 098, <www.camsys.com/pubs/NCHRP08-36-98.pdf>.

[38] For more information on Equal Employment Opportunity (EEO) Tabulation, see <www.census.gov/people/eeotabulation/>.

[39] The LODES data are primarily distributed in the OnTheMap Web tool available at <http://onthemap.ces.census.gov/>.

[40] The QWI data may be browsed using the QWI Explorer, available at <http://qwiexplorer.ces.census.gov/>.

Appendix: Design Difference Summary Table

Element	ACS	LODES
Collection	Survey of 3.5 million addresses in the United States and Puerto Rico per year collected continuously in monthly samples.	Administrative records from the unemployment insurance reporting systems of 50 states and the District of Columbia, and from the Office of Personnel Management.
Coverage	The universe of workers is limited to those 16 years and over who worked during the ACS reference week, in reference to when the respondent completes the questionnaire. Respondent may only report one job. In regards to coverage of workplace location, for ACS respondents who report working from home, the home is their primary workplace location, which differentiates them from workers who occasionally work from home or telework.	All jobs covered under state unemployment insurance law (95 percent of private sector wage and salary employment) plus most civilian federal employment. Does not cover the following groups: self-employment, military employment, the U.S. Postal Service, and informal employment. Job holders may be of any age.
Geographic and Longitudinal Scope	Decennial censuses 1960 through 2000. ACS tabulations with national coverage beginning with the 2000 ACS and full implementation starting with the 2005 ACS. Since the 2006 ACS, data for people in group quarters (e.g., college dormitories, military barracks) have been collected. Workers living in institutionalized group quarters are not included in the universe for commuting estimates.	The workplace domain includes states with available earnings data from 2002 onwards. The residence domain includes all states from 2002 onwards. From 2002–2008, job characteristics included ownership, industry, age, and earnings. From 2009 on, sex, race, ethnicity, and educational attainment are also available. Tabulations from 2010 on include federal workers. From 2011 on, firm age and firm size are also available. Changes in processing methodology and updates to data and imputations lead to some longitudinal inconsistencies.
Job Definition and Reference Period	Tabulations include those aged 16 and older who worked in the reference week, or the calendar week the questionnaire was completed, and who provided a discernable workplace location. Workers who were on vacation or sick leave the entire reference week are not included. Workers living in noninstitutional group quarters are included. Those living in institutional group quarters (e.g., prisons, nursing homes) are not included. ACS 1-, 3-, and 5-year estimates are period estimates, which mean they represent the characteristics of the population and housing over a specific data collection period. Data are combined to produce 12 months, 36 months, or 60 months of data.	For a job, or an earnings history of a worker at an employer, to be included in a year of LODES, there must be earnings in both the first and second quarter. Workers with earnings from an employer in both quarters are assumed to be employed at the seam of those quarters, or April 1, the first day of the second quarter. Thus, LODES constitutes a cross-section or snapshot of all jobs held at that reference date. This definition omits some very short duration jobs. Earnings categories and job dominance are based on second quarter earnings. The source residence information could refer to any point in the year.
Job and Worker Characteristics	Worker characteristics depend on tabulation. The standard tables include workplace location, commute mode, departure time from home, arrival time to work, travel time (minutes), sex, age, race, ethnicity, citizenship status, language spoken, earnings, poverty status, occupation, industry, class of worker, hours worked each week, weeks worked in the past 12 months, earnings, number of vehicles available, household size, number of workers in household. Tables can be rendered for both current residence and workplace locations.	Job characteristics include dominance (primary or secondary job), ownership type (private, all), NAICS industry sector, firm age and firm size, earnings, age, sex, race, ethnicity, and educational attainment. These characteristics are constructed from both the unemployment insurance wage records as well as linked administrative and survey records, and when missing, they are completed with imputation models.

Appendix: Design Difference Summary Table—Con.

Element	ACS	LODES
Location Definitions (Workplace and Residence)	The ACS questionnaire specifically asks for a work location of the respondent (employee) "last week," which may include a location associated with work-related travel that is not the respondent's typical work location if it occurred in the last week. The current place of residence is the housing unit or group quarter address being sampled. People at the address are included in the survey if they are living or staying at the location for more than 2 months or do not have another place to stay.	The employment location is reported by employers. In some cases this may not be the location at which an employee performs his/her work duties. Residence location is derived from annual federal administrative data. LODES includes no information on commute mode, or whether the origin-destination flow constitutes an actual trip.
Completeness of Geographic Information and Edits/Imputations	Among workers in the ACS sample, the workplace locations of about 92 percent of worker records are successfully coded to the place level, and the remaining cases are allocated a workplace location down to the place level.	For multiestablishment employers, establishments are not assigned to jobs in the source data except for in Minnesota. Candidate establishments allocated to jobs using multiple imputation based on establishment size and proximity to residence. LODES completes workplace geography by drawing from the distribution of workplaces of residential neighbors.
Geographic Tabulation Levels	Varies by release. ACS data are released as single-year data or multiyear pooled samples of 3 and 5 years. Across release types, ACS tables are restricted according to summary levels and population thresholds.	LODES job counts are released at the census block level. All higher-level aggregations produced by the OnTheMap application are sums of whole census blocks.
Control Totals	The ACS uses independently derived annual population estimates as population controls in its weighting methodology. The 2010 ACS estimates and later years use population estimates that reflect the results of the 2010 Census. Population controls are applied down to the county level for basic demographic characteristics: age, sex, race, and Hispanic origin. Population controls are applied down to the place and minor civil division level for basic population totals.	At the state level, the source data for LODES are reweighted to state totals from the Bureau of Labor Statistics (BLS) release of Quarterly Census of Employment and Wages (QCEW). Because of the small-cell imputation, noise infusion, some small-cell rounding effects, and input-data vintaging differences for the released LODES statistics, state totals may deviate slightly from QCEW totals, as well as from QWI.
Confidentiality Protection and Suppression	Some ACS data are suppressed in order to limit the disclosure of information about individuals or reduce the number of estimates with unacceptable levels of statistical reliability. Geographic areas or groups of 65,000 population or more are eligible for 1-, 3-, and 5-year estimates. Areas or groups of 20,000 or more are eligible for 3- and 5-year estimates. Areas or groups of 20,000 or fewer are eligible for 5-year estimates only.	Employment totals are protected by noise infusion and small cell imputation. Residential location is protected by synthetic data methods with probabilistic differential privacy. LODES releases protected job counts in all cells at all levels of aggregation, with no suppression.
Related Public-Use Data Products	American Community Survey (ACS) commuting data are distributed through the American FactFinder (AFF) Web site, with earlier years also available on the Census Bureau's FTP site. The Census Bureau produces county-to-county commuting flows as part of the routine update of metropolitan and micropolitan statistical area boundaries. The Census Bureau also provides Public Use Microdata Sample (PUMS) data from the ACS, and releases ACS-based reports on various commuting topics.	The LEHD Origin-Destination Employment Statistics (LODES) are distributed through the Web tool OnTheMap. The Longitudinal Employer-Household Dynamics (LEHD) program also releases the Quarterly Workforce Indicators (QWI), which provide more job measures, greater industry detail, and more longitudinal consistency, but less geographic detail.