Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files

by

Andrew S. Green
Cornell University & U.S. Census Bureau

Mark J. Kutzbach
U.S. Census Bureau

Lars Vilhuber
Cornell University & U.S. Census Bureau

CES 17-34 **April, 2017**

Abstract

Commuting flows and workplace employment data have a wide constituency of users including urban and regional planners, social science and transportation researchers, and businesses. The U.S. Census Bureau releases two, national data products that give the magnitude and characteristics of home to work flows. The American Community Survey (ACS) tabulates households' responses on employment, workplace, and commuting behavior. The Longitudinal Employer-Household Dynamics (LEHD) program tabulates administrative records on jobs in the LEHD Origin-Destination Employment Statistics (LODES). Design differences across the datasets lead to divergence in a comparable statistic: county-to-county aggregate commute flows. To understand differences in the public use data, this study compares ACS and LEHD source files, using identifying information and probabilistic matching to join person and job records. In our assessment, we compare commuting statistics for job frames linked on person, employment status, employer, and workplace and we identify person and job characteristics as well as design features of the data frames that explain aggregate differences. We find a lower rate of within-county commuting and farther commutes in LODES. We attribute these greater distances to differences in workplace reporting and to uncertainty of establishment assignments in LEHD for workers at multi-unit employers. Minor contributing factors include differences in residence location and ACS workplace edits. The results of this analysis and the data infrastructure developed will support further work to understand and enhance commuting statistics in both datasets.

Keyword*:* U.S. Census Bureau, LEHD, LODES, ACS, Employer-employee matched data, Commuting, Record linkage

---

*Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. This research was supported by the Improving Operational Efficiency (IOE) program at U.S. Census Bureau under the Dev10 project. Vilhuber acknowledges funding through NSF grant SES-1131848 (NCRN Cornell). We are thankful for comments from Martha Stinson, Erika McEntarfer, and Matthew Graham and from participants of the LEHD Research Workshop (2012), the NCRN Meeting (Spring 2014, Spring 2015), the 2015 Federal Committee on Statistical Methodology Research Conference, the 2016 Joint Statistical Meetings, and the Census Economic Research Brown Bag Seimnar (2016). The American Community Survey Office (ACSO) provided assistance in obtaining the necessary microdata extracts. The Social, Economic, and Housing Statistics Division (SEHSD) provided support for Dev10 and expertise on ACS processing, coding, and tabulation. LEHD staff provided assistance with data transfers, geocoding, and expertise on edits and imputations. The Summer Working-group on Employer List Linking (SWELL) developed the matching tool-kit used for this study and reviewed thousands of candidate matches.

1 Introduction

A wide array of users are interested in the U.S. Census Bureau's local employment data with joint workplace and residence characteristics, which they use in combination to make inferences on commuting. Transportation planning agencies and urban planners at the federal, state, and local levels use jobs and commuting data for measuring population density and for infrastructure planning [NCHRP, 2007]. The Office of Management and Budget (OMB) uses commuting flows to define metropolitan areas. OMB combines sets of counties with a high degree of social and economic integration with a core, as measured by commuting ties [Office of Management and Budget, 2013]. These areas are then used as input both for statistical purposes as well as for the allocation of federal funds. Workforce development agencies use the data to assess job availability.[1] Emergency response agencies use Census Bureau estimates of daytime population as well as linked home and work data for planning and to evaluate affected areas.[2] Businesses use data on employment, industry, and workforce concentration to make plant location decisions, with commuting as a consideration. Developers use data on the balance of jobs and housing to identify locations for commercial or residential construction. Researchers use the data to define Commuting Zones [Tolbert and Sizer, 1996], to evaluate theories of agglomeration [Fu and Ross, 2013] and spatial mismatch [Andersson et al., 2014], and to measure excess commuting [Horner and Schleith, 2012], to name just a few among many topics.

The Census Bureau produces statistics on commuting flows from two distinct data sources. The American Community Survey (ACS) records responses of employment and workplace location from a national, residence-based household survey. The ACS microdata is the input to several public use datasets with commuting information, including: county-to-county worker flows, the Census Transportation Planning Products (CTPP), and estimates of commuting behavior by home and workplace margins. The Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program assembles employer-employee matched administrative data with workplace and residence information for workers. The Census Bureau uses this jobs data to create the LEHD Origin-Destination Employment Statistics (LODES), a public use dataset providing residence-to-workplace flows, as well as other data products.

Given the availability of these two similar and widely used data sources, the purpose of the present analysis is to explain some of the differences in public use statistics. Early on in the development of LEHD, program planners, analysts, and state partners, examined commuting flows from survey and administrative data and found differences they attributed to unreported or incorrect establishment locations [Lane et al., 2003]. Users have noted that average commute distances and rates of between-county commuting tend to be higher in LODES than in ACS data products.[3] In an analysis of the public use data, Spear [2011] finds longer average commute distances in LODES

[1] For example, the State of California Employment Development Department explains the role of commuting in labor markets at `http://www.labormarketinfo.edd.ca.gov/data/county-to-county-commute-patterns.html`.

[2] See Commuter-Adjusted Daytime Population Data at `https://www.census.gov/hhes/commuting/data/daytimepop.html` and OnTheMap for Emergency Management at `http://onthemap.ces.census.gov/em/`.

[3] For a review of several case studies, see Murakami [2007].

compared with CTPP.

While the frames of the datasets and the information collected would be expected to have significant overlap, they differ in many respects (e.g. collection, coverage, definitions) [Graham et al., 2014] and neither is a travel diary or direct log of trips. ACS has many elements in common with the National Household Travel Survey (NHTS) [Federal Highway Administration, 2011a], asking about employer, workplace, and travel mode last week, but, unlike administrative records, requires self- or proxy response and allows only one job. LEHD assembles workplace, job, and residence information from separate sources. As such, the "origin-destination" flows published in LODES may differ from the typical travel survey concept. Henceforth, we use "commuting statistics" to refer to summaries of combined residence and workplace information from either source, even though differences in definitions will be a factor we consider for our explanations.

We believe the present analysis is the first to compare commuting responses from the Census Bureau's Journey to Work questions (in either the Decennial Census long form or ACS) with linked administrative records and that it is also the first to compare LEHD home to work flows with linked survey records.[4] In evaluating sources of discrepancies, we focus on commute distance and within-county commute rates - two economically meaningful measures that are sensitive to differences in the workplace and residence locations of workers. Distance is calculated point to point and within-county commute rate gives the share of persons who work in the same county where they live. By imposing various sample and definitional restrictions, this study traces these commuting statistics from public use data through to common sets of persons and jobs linked at a microdata level.

Figure 1 presents the difference that our linked microdata analysis will resolve. In the figure, each horizontal bar gives the within-county commute rate (a percentage) for a different sample. The bars labeled "LEHD" and "ACS" signify that the statistic is computed with both residence and workplace locations from the respective source file. Our analysis attempts to explain the longer commutes in LODES relative to ACS, which, even though counties vary widely in size, are evident from the lower rate of within-county commuting in LODES (54.9 percent) relative to ACS (72.5 percent). By linking the LEHD and ACS microdata at a person and job level, we produce the ACS-LEHD *Employer Match* sample, which has a within-county commute rate of 54.7 percent using LEHD home and workplace locations. The similarity of commuting statistics for this matched sample to the LODES public use data suggests that differences in person or job frames are not responsible for the disagreement.

In the subsequent analysis, we consider additional intermediate steps between the bars labeled (3) and (8) and reveal factors that are important for closing the gap. One major factor relates to differences in workplace frames between LEHD and ACS responses. The remainder is mostly

[4]Graham and Ong [2007] link job holders from the 2000 Census long form with LEHD to examine commuting patterns of single and dual job holders (as well as other topics in a series of related papers). They calculate distributions of commute time (from the survey) and commute distance (from LEHD) for different subsamples, but do not directly compare commuting for linked persons. Hyatt [2015] measures the correspondence of workplace location for married and unmarried partner households in the 2000 Census long form and LEHD, but the study is for a subset of workers and focuses on relative workplace locations rather than commute distance. [Isenberg et al., 2013] compared linked worker samples of ACS and LEHD in terms of industry.

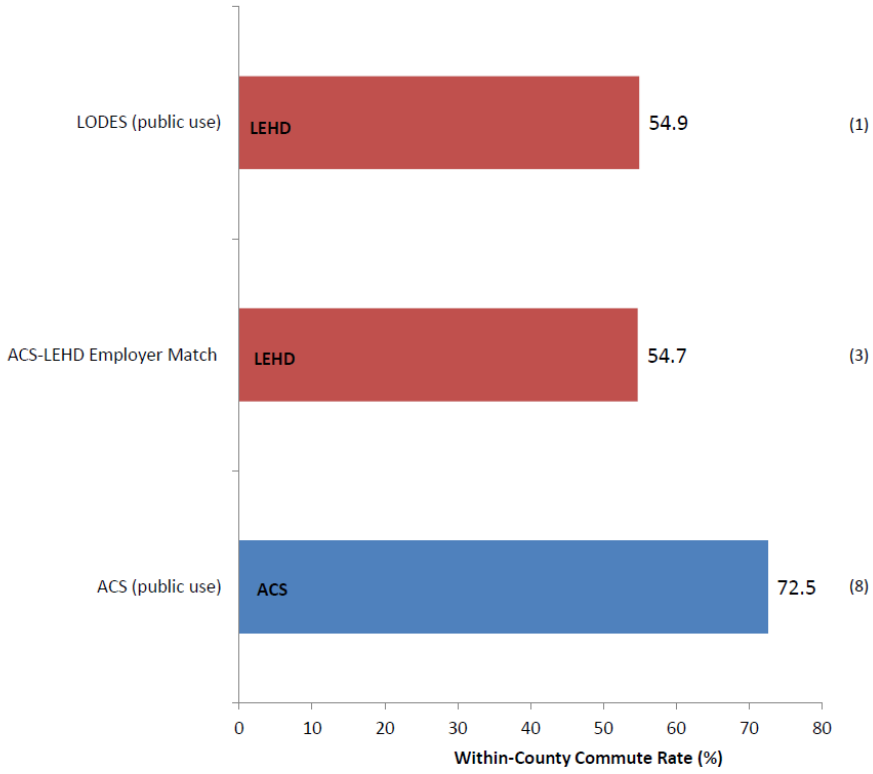


Figure 1: Summary of Within-County Commute Rate by Jobs Sample

Notes: Bars denote within-county commute rate for each sample, with red bars for LEHD home and workplace locations and blue bars for ACS home and workplace locations. Row numbers refer to the more detailed version, Figure 12.

explained by uncertainty in the assignment of establishments to workers for multi-unit employers in LEHD.

The present analysis does not provide a methodology for data users to adjust specific queries to be consistent across both datasets. Differences in the public use data for any particular query may depend on local circumstances for each dataset as much as the overall findings reported here. Nevertheless, for many broad-based analyses, the findings here should help to explain differences in prima-facie similar statistics computed for ACS and LODES. These findings of this study, and the data linking infrastructure developed for this project, will guide future efforts to enhance and explain public use data products.

Our study proceeds as follows. Section 2 reviews the ACS and LEHD source data and compares the public use data products. Section 3 describes the matching procedure and evaluates the characteristics of the matched sample. Section 4 compares commuting statistics for the matched survey and administrative data. Section 5 reviews our findings and Section 6 concludes by describing

potential further uses of the matched sample for improving the quality of public use data products.

2 Background

2.1 Overview of ACS and LEHD and Public-Use Commuting Statistics

Graham et al. [2014] conduct a design comparison of ACS and LEHD in terms of collection, coverage, geographic and longitudinal scope, job definition and reference period, job and worker characteristics, location definitions, completeness of geographic information, geographic tabulation levels, control totals, and confidentiality protection. We summarize several of these elements here inasmuch as they relate to commuting statistics.

The American Community Survey (ACS) is the successor of Decennial Census long form surveys that have recorded information on workplace location and commuting since the 1960 Census.[5] However, unlike the 2000 Census long form, which was a one-in-six household sample, the ACS is approximately a one-in-fifty household sample each year, and thus not sufficiently large to provide neighborhood level commute flow information for a single year. Rather, commuting flows and neighborhood level statistics are calculated from a pooled 5-year file that comes closer to the long form sample size. In the field since 2003, and with an expanded sample since 2005, the ACS in 2010 was based on survey responses from a mailing frame of about 3 million residences a year and about 2 million interviews. Data are collected continuously throughout the year (as opposed to the April 1 reference date for the Decennial Census). Respondents provide demographic and housing information as well as (for those age 16 and over) employment status and an employer's industry, name, and address. Based on the latter, geocoding is used to assign tabulation geography (e.g. census block, county) to a workplace. In addition, respondents report job information including their occupation, hours, commute duration (in minutes), time of departure, and commute mode (e.g. car, bus, walked, worked at home). The Census Bureau releases commuting and place of work information in several formats.[6] First, "Journey to Work" provides estimates of worker flows between counties, labeled as "County to County Commuting Flows for the United States and Puerto Rico: 2009-2013," as well as other statistics and trends [U.S. Census Bureau, 2015].[7] These county-to-county flows are an input to the delineation of metropolitan areas by OMB. Second, ACS tabulations by residential geography provide estimates of commuting behavior, such as the share of workers in a Census tract who travel by bus. Third, more detailed flow tables are available in the CTPP, which was also produced for the 1990 Census and the 2000 Census.[8]

The LEHD program at the Census Bureau produces LODES on an annual basis beginning with data from 2002. LEHD is a jobs frame, consisting of employer-employee matched administrative

[5] For an analysis comparing commuting data from long-form Census responses with the ACS, see McKenzie [2015].

[6] Extensive documentation on the ACS can be found at `https://www.census.gov/programs-surveys/acs`.

[7] The ACS tabulation, as well as other aggregations, are available at `http://www.census.gov/hhes/commuting/data/`.

[8] The American Association of State Highway and Transportation Officials (AASHTO), a non-profit association, partners with the Census Bureau to produce the CTPP, available at `http://ctpp.transportation.org/Pages/5-Year-Data.aspx`.

data provided by states and the federal government [Abowd et al., 2009]. States provide LEHD with quarterly files listing the earnings of all jobs covered by Unemployment Insurance (UI) as well as files giving employer account information including location, industry, ownership, and size. The employer account reports, also known as the ES-202 program, include the same information as states provide the Bureau of Labor Statistics (BLS) for the Quarterly Census of Employment and Wages (QCEW). These data are augmented with records on federal workers to provide a frame that covers government jobs as well as 96 percent of the private sector workforce [Stevens, 2007]. In 2010, LEHD included approximately 130 million jobs, held by 120 million workers, linked to 7.6 million establishments at 6.2 million employers. LEHD combines the employer-employee frame with federal administrative data on place of residence as well as with survey and administrative data on worker characteristics, including age, sex, race, ethnicity, and educational attainment. The program uses the combined data, known as the LEHD Infrastructure Files, to produce several public use data products.[9] LODES tabulates job counts with employer and worker characteristics by workplace, residence, and origin-destination margins at the Census block level (as well as more aggregate geographies) for jobs held on April 1 of each year (the beginning of the second quarter).[10]

2.2 How Design Differences May Contribute to Differences in Commuting Statistics

While we expect agreement in many respects, design differences between ACS and LODES may lead to some differences in public use statistics. Graham et al. [2014] provide a qualitative assessment of these differences, including person and job frame differences, the worker versus employer perspective, the handling of missing data, and confidentiality protection measures. We review these here as motivation for our empirical analysis.

ACS and LODES present different refinements of the universe of working persons, which will carry through to statistics on workplace and commuting. ACS tabulates responses from a frame of households at known addresses, weighted to the estimated population totals. LODES tabulates administrative data on UI covered jobs, weighted by workplace state to employment totals released for the QCEW. Putting aside issues of the job definition for these tabulations, the set of persons covered by these files may not completely overlap and the weights are for distinct baselines. Though both programs are mandatory and compliance is high, unit non-response by households and "UI holes" in LEHD would reduce representativeness.

Among persons accounted for in both databases, job definitions will affect tabulations as well as commuting statistics. Whereas ACS commuting flow tabulations include all those who say they are employed, the source of LEHD earnings records does not cover self-employed workers

[9] In addition to LODES, the LEHD program produces the Quarterly Workforce Indicators, which describe employment dynamics for the nation as well as states, counties, metropolitan area, and Workforce Investment Board areas. LEHD also produces Job-to-Job flows, which describes transitions of workers between employers, industries, and states, and into and out of employment. For more information on these data products, see `http://lehd.ces.census.gov/data/`.

[10] For the LODES data and technical documentation as well as the OnTheMap web tool, see `http://lehd.ces.census.gov/applications/help/onthemap.html`.

(also not covered are the armed forces and certain federal agencies, postal workers, international organizations, some non-profits, and workers in some family businesses - see Stevens [2007]). ACS respondents may under-report short term jobs [Abraham et al., 2013] and are unable to report a second covered job because ACS only collects one employment response.

Workers may report based on a different perspective in survey data than employers do in administrative data. The ACS surveys workers where they live, asks them whether and where they worked *last week*, and asks how they commuted to that location.[11] In most cases, these responses (where they worked the most hours last week) should be a worker's regular worksite. However, the location reported might not be the *usual* workplace if, for instance, the worker attended a client or conference in another city. The ACS does not have a follow-up question to confirm whether the reported location is the usual destination. When a respondent reports working from home, ACS tabulates the reported workplace to conform to the place of residence. An ACS responder's residence is the sampled location from the Master Address File to which the Census Bureau sends a form, invitation, or interviewer.

While employers in LEHD are supposed to report all establishments where workers perform their duties or where they are supervised from, in some cases, those locations may be places that a worker never, or only occasionally, visits. Thus, a worker appearing in administrative data at one workplace may not report the same location on a survey, though both locations may be informative about an employment relationship. LEHD derives residence location from a variety of federal records that are combined and de-duplicated by person-year.[12] At the reference date for a job (April 1, for LODES), the LEHD residence location may differ from a worker's home if that worker has a different address in federal records, or moves during the year.

Missing or incomplete data are a challenge for both programs. Responses to the ACS are subject to item non-response as well as reporting and recording errors. In 2010, 92.4 percent of ACS workplace responses could be geocoded for county level assignment.[13] When targeting more detailed geocodes, such as census blocks, the geocoding success rate is lower (discussed below). ACS imputes missing workplaces using a hot-deck model based on the responses of neighbors.[14] LEHD is able to geocode over 97 percent of reported workplace locations to a Census block level,

[11]The ACS Questionaire for 2010 includes the following, relevant questions, used in this report: 29a. "LAST WEEK, did this person work for pay at a job (or business)?" 29b. "LAST WEEK, did this person do ANY work for pay, even for as little as one hour?" Responses to these questions are combine to determine employment status. Workplace location is based on the question: 30. "At what location did this person work LAST WEEK?" For a person's current or most recent job activity, or the one with the most hours, the survey asks about the type of job (41.) and employer name (42.), among other questions (e.g. industry, occupation).

[12]For a discussion of the original Composite Person Record and Statistical Administrative Records System (StARS) methodology see Vilhuber and McKinney [2014], Abowd et al. [2009]. For the LEHD residence methodology from 2011 onward, see Graham et al. [2016].

[13]Authors' calculation from American FactFinder, Table B99081, IMPUTATION OF PLACE OF WORK, Universe: Workers 16 years and over, 2010 American Community Survey 1-Year Estimates.

[14]An automated geocoder assigns geography to 53 percent of place of work responses. The unassigned are reviewed by computer-assisted clerical coding operators. A hot deck procedure imputes geography for remaining records with sorting variables including industry groupings, means of transportation to work, minutes to work, state of residence, county of residence, and the state in which the person works [U.S. Census Bureau, 2009].

and LODES imputes the remainder using the distribution of neighbors' workplaces.[15]

An issue unique to the LEHD is that employers may under-report establishment locations, and when they do report them, most states do not require assigning establishments to jobs in the UI earnings records. Employers with multiple workplaces are required to complete a Multiple Worksite Report (MWR). Although compliance is not perfect, Spear [2011] reports that 97% of all covered employment is represented by reported establishments.[16] LEHD uses imputations to fill short term gaps in worksite reporting, but longer term under-reporting will tend to concentrate employment at the location of the employer's headquarters. Approximately 44 percent of all jobs in the LEHD Infrastructure files are at multi-unit reporting employers.[17] In order to assign establishments to workers, LEHD uses an imputation model, known as the Unit-to-Worker (U2W) imputation [Abowd et al., 2009, Stephens, 2007], which attempts to replicate the distribution of establishment sizes at a firm as well as the known distribution of commute distances overall. Thus, while the model favors larger and closer establishments, workers can also be assigned a location far from home. LEHD data products represent the uncertainty of assignment by including the equally weighted results of 10 independent draws from the imputation model.

To illustrate how establishment assignment weighting works, consider the following three examples:[18]

1. A worker's employer only has a single location (a single-unit employer). For single-unit employers, the same workplace is included with unity weight in all tabulations.

2. A worker's employer has two, equal sized, establishments (the smallest possible multi-unit employer), with workplace A being 1 mile from the work er's residence, and workplace B being 20 miles from the worker's residence. The imputation model favors A (it is the more likely workplace), but the 10 draws yield A 6-times, and B 4-times, each of the draws being given a weight of 0.1. Thus, the de-facto weight on A is 0.6, and the weight on B is 0.4.

3. A worker at a multi-unit employer with more than 10, separately located establishments, could receive draws from the imputation model at up to 10 different locations. In compiling employment statistics, the job record would be allocated with a weight of 0.1 to each of the 10 locations (with remaining establishments given zero weight).

Uncertainty in model input data may add noise to origin-destination flow data, which the multiple imputation approach is designed to accommodate. Restrictions to the candidate list of establishments due to under or misreporting of establishment locations, and mis-specification of the imputation model may add bias to origin-destination flow data, which the multiple imputation

[15] LEHD processes address data through the Geocoded Address List (GAL) process, which deduplicates addresses (combining those that are equivalent) using a commercial geocoder and assigns a unique identifier along with precision metrics to each deduplicated address. Given the compliance obligations within UI systems for providing accurate information on establishments, employers-provided addresses typically conform to standard mailing address formats.

[16] For the BLS summary of MWR as well as links to forms for each state, see `https://www.bls.gov/cew/cewmwr00.htm`.

[17] Authors' calculations using LEHD data. Spear [2011] reports 45% based on BLS internal statistical analysis.

[18] We thank Martha Stinson for suggesting these examples.

Table 1: Comparison of ACS, LODES, and LEHD

Characteristic	ACS (public use) 2009-2013	LODES (public use) 2011	LEHD Infrastructure Files 2011
Total jobs (millions)	139.7	120.3	119.5
Total non-zero flows (out of 9,790,641)	137,492	614,291	978,321
Within-county commute rate	0.725	0.549	0.553
Average commute distance (miles)	N.A.	31.1	50.0

Note: ACS 5-Year statistics constructed from U.S. Census Bureau [2015]. LODES and LEHD Infrastructure Files statistics are for primary jobs held on April 1, 2011. For computational details, see text.

approach does not protect against. For instance, under-reporting of establishments that happen to be nearby a workers home (and likely to be selected by U2W) could lead to a longer commute for that job. In aggregate, such instances could increase average commute distance.

A further issue for comparisons is the difference in the geographic and longitudinal frames of the data. Survey-based data on county-to-county commuting flows are available for the 2000 Census and from 2005 onward with ACS. Unlike the 2000 Census, commuting flows form ACS require a 5-year pooled sample for complete coverage. While LODES is available annually since 2002, it does not cover all states in all years. The Local Employment Dynamics (LED) partnership is a voluntary agreement of states and the Office of Personnel Management (OPM) with the Census Bureau to produce LEHD. While all states have joined the LED partnership at some point, not all states are available in the public use data in all years. The only years of LODES including all states are 2011-2013. LODES also does not include Puerto Rico or international workplace destinations, which are reported in the ACS.

The Census Bureau uses confidentiality protection measures to ensure that jobs and commuting data represent general patterns, not the information of particular individuals. We briefly describe these measures here and present some relevant comparisons below, but leave deeper analysis for later work. The ACS uses a methodology referred to as "swapping" to exchange the information of similar records, which adds a degree of uncertainty to commuting statistics [U.S. Census Bureau, 2009]. Furthermore, the ACS suppresses statistics in cells not meeting publication requirements, as is documented for commuting statistics in Spear [2011]. LEHD infuses multiplicative noise [Evans et al., 1998] to employer job counts [Abowd et al., 2009] and uses synthetic data techniques to provide probabilistic differential privacy for the residential locations associated with jobs [Machanavajjhala et al., 2008]. Both methods result in lower quality information at very detailed levels but retain high quality information at more aggregate levels.

2.3 Public-use Data Comparison

Table 1 provides a summary of commuting flows aggregated to the county level for ACS, LODES, and the LEHD Infrastructure File (henceforth, LEHD, for this subsection). ACS tabulations are derived from U.S. Census Bureau [2015]. For LODES and LEHD, we use an April 1, 2011 tabulation, consisting of private sector as well as state, local, and federal government workers. The LODES/LEHD year is at the mid-point of the pooled ACS file. LODES statistics are produced from the published data underlying OnTheMap, which have various edits as well as comprehensive disclosure avoidance applied. LEHD statistics are produced directly from the confidential microdata. Because, ACS only reports one job per respondent, we limit LODES and LEHD to primary jobs, or the highest earning job for each worker, to be comparable to ACS.[19]

Table 1 shows that within-county commute rates in LODES are longer than the comparable statistics derived from ACS. First, note from the first row in Table 1 that ACS flows data represent substantially more jobs than LODES (LODES has slightly more jobs than LEHD because of rounding applied in the aggregation). This discrepancy is due to the coverage difference of the jobs frame and almost entirely explained by the self-employed in ACS [Spear, 2011]. However, LODES, with 614,291 origin-destination flows between counties, has more than four times as many observed pairs as ACS (LEHD has even more possible flows because it retains more information on the uncertainty of workplace locations).

While not a direct measure of distance, the share of workers commuting to jobs in the same county captures the tendency to commute to jobs closer to home. We compute the average of within-county commute rates across all counties, weighting by the count of workers residing in each county in the respective datasets. Thus, these statistics, as with the remainder of the report, are job-weighted rather than being weighted by tabulation cell (county, in this case). While only 54.9 percent of workers in LODES work in the same county where they live, as do 55.3 percent in the LEHD Infrastructure Files, 72.5 percent do so in ACS. Commutes to nearby workplaces are relatively more common in ACS.[20]

The ACS county-to-county flows do not have sufficient geographic detail to compute average commute distances for typical commutes, but we are able to do so for LODES and LEHD. The average commute distance of 50 miles for LEHD falls to 31.1 miles in LODES.[21] We have found that the drop in commute distance is not broad-based, rather, it is concentrated as a reduction in very long distance commutes. We believe this drop is a result of features of the confidentiality protection system that were not tuned to preserve rare, long distance commutes. Despite this difference in

[19]The LEHD Infrastructure Files summary is produced from the WHATB, an intermediate file in the production of LODES. Both the LODES and LEHD summaries use the QWI protections for job counts, but only LODES uses the synthetic data technique to protect residence location. For this summary, we rounded the LEHD flows to the nearest integer and dropped flows that rounded to zero, reducing the total count of flows.

[20]AASHTO [2015] reports a corresponding value for a similar tabulation of the same ACS data product, and also shows that the rate of outside-of-county commuting has almost doubled since 1960.

[21]The CTPP, which has greater geographic detail, would provide a means to compute commute distance for an ACS-based data product. As an alternative reference point, the NHTS reports an average commuting distance of 11.8 miles for 2009 [Federal Highway Administration, 2011b].

commute distance, the within-county commute rate is very similar for LEHD and LODES. In our analysis, we focus on within-county commute rate for comparisons with the public use data, but consider both that measure and distance when investigating the linked microdata sample.

To provide a geographic perspective on ACS and LODES commuting data, we map the within-county commute shares in Figure 2. Consistent with the statistics in Table 1, the ACS map in Figure A4a is generally darker than the LODES map in Figure A4b, indicating more within-county commuting according to the ACS. The differences between the two maps are pervasive across all regions. We include more detailed maps of selected states in the Northeast, Midwest, South, and West in the Appendix.

The maps highlight several factors in commute distances that this study will not focus on that influence both ACS and LODES statistics. First, with regard to using county flows as a measure of commuting, the fact that spatially larger counties have fewer outflows is not surprising. Second, counties in geographically isolated areas (such as southern Florida), inaccessible areas (such as Appalachia and the Rocky Mountains), or both (northern Maine) have greater local flows. Third, urban centers have larger shares, but the suburban counties surrounding them tend to have lower shares. For example, Minneapolis, Minnesota (Hennepin County) is shaded darker than surrounding counties, consistent with commuting flows to the central business district.

3 Data Linkage

This study makes use of several extracts from confidential microdata on a secure server. We use the pre-swapped, edited versions of the ACS, linked with write-in files for 2009 and 2010.[22] The person and write-in files are linked by household and person number fields (CMID and PNUM). We also use these fields to crosswalk person records with unique person IDs, or PIKs, described below. The LEHD Infrastructure Files are updated quarterly using the latest code and inputs. For this study, we made extracts in 2014 for jobs held anytime from 2004 through 2011. The extracts do not include jobs provided by Massachusetts or by OPM, which were not fully integrated at that time and not available for the entire study period.

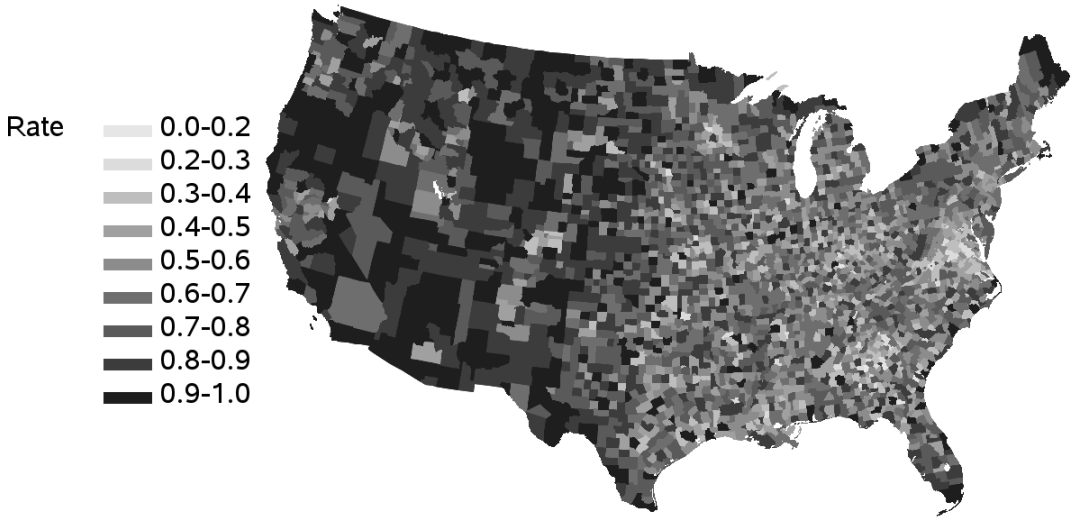
3.1 Blocking Strategy for Narrowing the Candidate Set

The ACS-LEHD linkage project began in 2010 as a collaboration between the Center for Economic Studies and the Social, Economic, and Housing Statistics Division.[23] The project's goal is to integrate these two data sets at the job level in order to examine inconsistencies between survey and administrative data on, in the case of this study, place of work and commuting information.

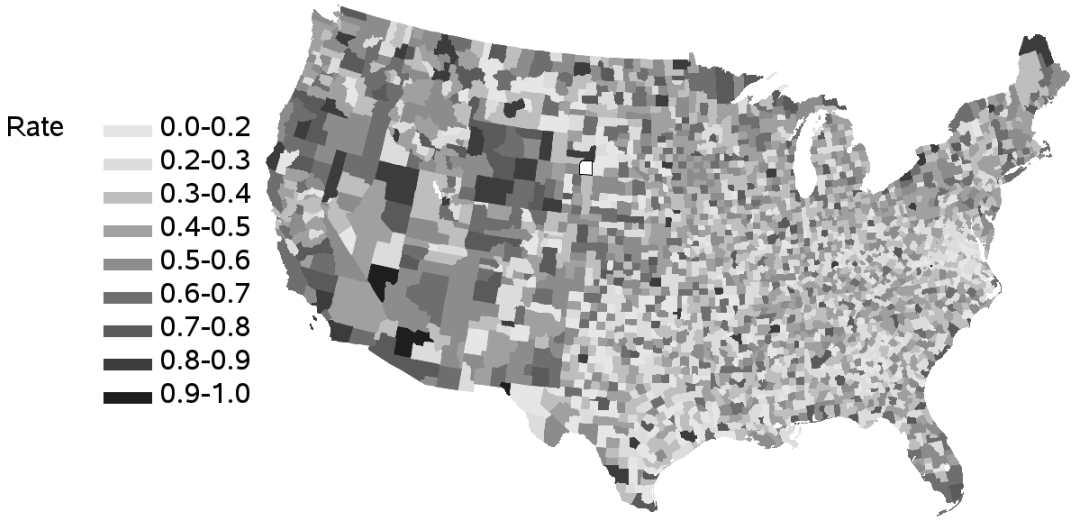
Data integration takes places at two levels. Because the LEHD data is hierarchical, with establishments belonging to employers, we link ACS responses both to a job - defined as an employer

[22] ACSO provided these files for 2005 to 2010 for the this project. Our research integration is only for the last two years, but the same methodology could be applied over a longer time span.

[23] The linking project is known as Dev10 at the Census Bureau and was supported by the Improving Operational Efficiency program.



(a) ACS



(b) LODES

Figure 2: Within-County of Residence Commute Rate

Notes: Shading coresponds to a larger share of county residents commuting to a workplace in the same county where they live, by decile bins from 0 to 1. See Table 1 for definitions of ACS and LODES public-use data.

within a state - and to workplaces at that job. In principle, all establishments and employers could be considered candidates for a match to ACS. However, the number of possible pairs generated from combining 1.5 million ACS employment responses in a year with 7.6 million establishments at 6.2 million employers is enormous. In order to reduce the set of possible linkages, we leverage personal identifying information provided by ACS respondents.[24]

Table 2: Sample Restrictions for Analysis Sample

Sample Restriction	Person Records	Percent
ACS respondents in 2009 and 2010	$\approx$ 9,000,000	
Age $\geq$ 16 at response date with PIK assigned (92%)	6,666,000	
10% random person sample	667,000	10.0%
ACS employed	**311,000**	46.6%
Linked to any LEHD job in 3-quarter window around response date	287,000	92.2%
Sufficient ACS information for matching (name, address)	285,000	91.6%
Sufficient LEHD information for matching (name, address)	277,000	89.1%

Notes: Records rounded to 1,000s. Percentages computed relative to the line immediately above.

Table 2 provides observation counts for the restrictions and blocking process. From the approximately 9 million responses in 2009 and 2010, we limit our analysis to those age 16 or older at the response date, which is the minimum for an employment response to be recorded. The first step in linking ACS employment responses to LEHD job candidates is to assign unique identifiers to persons in each dataset. The Census Bureau assigns Protected Identification Keys (PIK) to ACS person records based on agreement of administrative records with survey responses (name, date of birth, sex, and place of residence) [Wagner and Layne, 2014]. Approximately 92 percent of ACS person records can be linked to a PIK.[25] After restricting on age and the PIK requirement, we limit the analysis to a 10 percent random sample, leaving us with 667,000 person records.

The next step is defining a frame of jobs in each dataset that potentially overlaps (though non-overlap may contribute to differences in commuting statistics, an issue we will examine later). The ACS asks respondents about the job they held in the week prior to the date of the survey. We define a respondent as *ACS employed* if she worked "last week" (`WRK=1`) and was coded as "employed, at work" (`ESR=1`). In addition, in order to make the ACS employment definition more comparable to LEHD coverage, we require that her dominant job was likely UI-covered. This definition corresponds to the restriction that the ACS class-of-worker variable `COW` $\in \{1, 2, 3, 4\}$.[26]

[24] An alternative would be to narrow the candidate set by geography, time frame, or perhaps industry agreement with ACS, but such parameters would bias matches to have a high degree of agreement on job characteristics - the main focus of our comparison.

[25] In this study, we do not explore the sensitivity of the observed comparison to the PIK assignment process. In the infrequent event of multiple persons being assigned the same PIK, we randomly select only one record.

[26] The `COW` codes are: (1) employee of a private for-profit, (2) employee of a private not-for-profit, (3) a local government employee, and (4) a state government employee. Federal government employees are not covered by

About 46.6% of eligible ACS respondents satisfy this criterion (see Table 2). Our requirements are more restrictive than the standard ACS employment definition, which reports 57.0% of eligible persons employed in 2010.[27] Our ACS-employed sample consists of 311,000 workers.

To block employment responses with LEHD jobs, we require an exact match on person and an approximate match on the timing of the job. LEHD job records consist of a PIK, identifying a person, a State Employer Identification Number (SEIN), identifying an employer within a state, and a sequence of quarterly earnings records.[28] The Census Bureau assigns a PIK to LEHD job records based on a crosswalk from employer-provided SSNs.

For the approximate match on timing of the job, we restrict the candidate set of LEHD jobs to those with earnings in either the ACS response quarter, or in the two adjacent quarters. This window reduces the candidate set from all recorded jobs to only those active around the response time, but accommodates a potential lack of precision in the timing of LEHD earnings records. Earnings records aggregate all pay periods ending within a three-month quarterly period, with no distinction of when in the quarter the job was held. While this window may add some candidate jobs not worked "last week," it accommodates responses at the beginning and end of the quarter.[29] In the case of employer identifier and name changes in LEHD, this window also allows for more flexibility in matching by including candidates for both the preceding and succeeding name, only one of which may match an ACS response.[30]

With this blocking, we link 287,000, or 92%, of our *ACS employed* sample to an LEHD job. From these persons, all combinations of persons with establishments at these jobs resulted in a linked file of over 15 million records. We performed the matching analysis described below on this set of candidate pairs.

3.2 Methodology for Matching Responses to Jobs and Workplaces

Our matching approach has features that both facilitate the linkage process and enhance the quality of the linked data that result.[31] First, the blocking described above significantly reduces the

state unemployment insurance systems, and are excluded here. Note that federal employment tracked through OPM has been added to LEHD and is included in LODES, but was not included in this research extract. While there are efforts to produce tabulations of self-employment from administrative data, these were also not included in the research extract.

[27] See American FactFinder, Table S2301, EMPLOYMENT STATUS, Universe: Population 16 years and over, 2010 American Community Survey 1-Year Estimates.

[28] SEIN is a LEHD-generated identifier, using information provided by state agencies. It is distinct from the "Employer Identification Number" assigned to employers by the IRS. In particular, it only identifies an employer within state boundaries.

[29] For example, a new employee responding to the ACS in the last week of a quarter may say that he is employed, but if the employer's pay period carries over into the next quarter, he would not appear in the administrative data in the quarter of hiring.

[30] The establishment candidate set is the list of establishments that were active (had positive employment) at an employer in the ACS response quarter if the worker had earnings in that quarter. If the worker only had earnings in the first or third quarter of the window, we only consider establishments that were active in those quarters, proceeding in the order as described (i.e. response quarter, first quarter, third quarter).

[31] This study makes use of job matching techniques developed by the Summer Working-group for Employer List Linking (SWELL). This collaborative effort between researchers at the Census Bureau, the University of Michigan, and Cornell University developed a toolkit for use in several linking projects [Gathright et al., 2016].

candidate set. Second, we implement new standardizing techniques that take into account features that are common in business names as well as a high quality address standardizer. Third, we create and leverage a human-reviewed set of candidate pairs to use in training our matching models.

The matcher makes use of employer name and (geocoded) address information provided by the respondent in the ACS and by the employer in LEHD. String comparators are used for name fields, whereas the latitude-longitude information from the address can be used to compute measures of proximity. The LEHD data include up to three name fields for each establishment - a legal, trade, and worksite name - and may include both a physical and mailing address. For matching, we make use of all of the name fields and prefer the physical address when available.[32] We use LEHD geocoding (described in section 2.2) for both ACS responses and LEHD employers, ensuring that no differences arise due to variation in the geocoding process. We use the point coordinates obtained to calculate the distance between ACS and LEHD workplace addresses and use this measure to construct a log-scale comparator of proximity (or spatial agreement). Table 2 shows that 277,000 records, or 89.1% of those linked to a candidate, have sufficient detail to allow for matching, but a small but significant fraction on both ACS and LEHD lack such information.

Researchers reviewed over 3,000 ACS-LEHD person/employer/establishment records in constructing the clerical review set for training the matching model.[33] Using this training data, we estimate logistic models explaining employer and establishment match status with string and spatial comparators for name and address agreement.[34] Using the parameter estimates, we predict employer and establishment match probabilities on the complete set of candidates. We designate a minimum predictive score threshold for matches based on a 5% false match rate (calculated using a reserve portion of the truth set). Any person/employer or person/establishment pair with a predicted probability above the threshold is then deemed a match.[35]

3.3 Employer and Establishment Matched Samples

Table 3 provides an overview of the results from the matching exercise. Of the 311,000 ACS respondents employed in likely UI-covered jobs (see Table 2), we match 226,000, or 72.7 percent (81.6 percent of the 277,000 with sufficient information), to at least one LEHD employer, and 114,000, or 36.7 percent (41.2 percent of the 277,000 with sufficient information) to at least one establishment at these employers. We refer to these subsets as the *Employer Match* sample and the *Establishment Match* samples, respectively. For the Employer Match sample, persons are matched to, on average, 1.035 jobs and have 19.791 candidate establishments. While name agreement was relatively more

[32]We use an employer name standardizer developed in collaboration with SWELL [Wasi and Flaaen, 2015] and a Jaro-Winkler string comparator to measure the similarity between the ACS name response and the closest LEHD match.

[33]Reviewers from the SWELL team include Graton Gathright, Kristin McCue, Holly Monti, Ann Rodgers, Kelly Trageser, Nada Wasi, and Christopher Wignall, in addition to the authors of this article.

[34]We considered both logistic and Fellegi-Sunter matching models. Because we achieved higher overall match rates with the logistic model for the same false match rate, we only report the results based on the logit predictions and do not further discuss the Fellegi-Sunter matching from the results.

[35]In the rare cases where a person/establishment record matches where a person/employer record failed, we also designate the person/employer record to be a match.

Table 3: Summary of Matching Results for ACS Employed Sample

Sample Restriction	Person Records	Percent
ACS Employed	311,000	
Employer Match		
Match to any LEHD Employer Candidate	226,000	72.7% (of ACS Employed)
Distance restricted - commutes less than 200 Miles	158,000	69.9%
Establishment Match		
Match to any LEHD establishment candidate	114,000	36.7% (of ACS Employed)
Distance restricted - commutes less than 200 Miles	92,000	80.7%

Notes: Percentages are computed relative to the line immediately above, unless otherwise noted. For details on the ACS employed sample, see Table 2. Sample restriction to commutes less than 200 miles is all ACS respondents who have commutes less than 200 miles and all corresponding LEHD employer matches have at least one establishment with a commute less than 200 miles.

important for employer matching, address agreement was crucial for establishment matching. From the validation analysis of our truth set, we approximately achieved our targeted false match rate of 5 percent. In examining some remaining cases of non-matching, we generally confirmed that no candidate was appropriate, suggesting that even with the class of worker restrictions in ACS, some job frame differences remain.

Appendix Tables A2 and A3 break down the match rates by ACS characteristics and ACS industry, respectively. Employer match rates are substantially lower, approximately 33 percent, for those not responding by mail (using CATI/CAPI). Industries where workers are typically required to report to a regular worksite tend to have higher establishment match rates. For example, match rates are higher for workers who report in the ACS that their employer is in the manufacturing industry (42.3 percent), where facilities are large, long lasting, and require the presence of workers. Match rates are lower in construction (24.0 percent), where an employer may not consider a temporary worksite to be an establishment. One exception is public administration, where establishment match rates are lower (29.1 percent) because many state and local governments do not report multiple worksites, such as schools within a school district.

The last step in Table 2 is our restriction to a shorter commute sample. We define "distance restricted" subsets, limited to ACS responses who (1) have ACS commutes of less than 200 miles *and* (2) have at least one candidate LEHD establishment that is less than 200 miles from their LEHD-reported residence. We believe 200 miles is a reasonable upper bound for an American daily commute, though the restriction does not imply that longer distance home-to-work flows are invalid. The *distance restricted Employer Match* and *distance restricted Establishment Match* samples contain 158,000 and 92,000 persons, respectively.

4 Analysis

4.1 Commuting Statistics for Comparison Analysis

We compare commuting across samples using two statistics: *average commute distance* (in miles) and *within-county commute rate* (sometimes presented as a percentage). Commute distance between a residence and workplace is the Great Circle distance, in miles, from one set of latitude and longitude coordinates to the other. Within-county commute status is an indicator $\in \{0, 1\}$, set to unity when the residence and workplace locations are both within the same county and zero otherwise. The within-county commute rate tends to fall as average commute distance rises. Without loss of generality, we will use d_i^S to refer to either of the measures for person i, but use "distance" for clarity of exposition. Superscript $S \in \{A, L\}$ denotes the source of the measure, either ACS or LEHD.

The commuting statistics, d^A and d^L, are weighted averages for a sample population N computed as $d^S = (1/N) \sum_{i=1}^{N} d_i^S w_i^S$. The weighting term, w_i^S, is calculated to make the sample statistics representative of ACS or LODES, the respective public-use datasets. Weighting to public use data is based on worker and job characteristics (including age, sex, and industry sector).

We compare commute distance under several different scenarios. ACS distances, d_i^A, are straightforward, because each respondent has only one residence and one workplace.[36] LEHD distances require aggregation across jobs and workplaces to the person level because there may be multiple employers with multiple establishments. Formally,

$$d_i^L = \sum_j p_{i,j} w_{i,j}^L \left(\sum_e p_{i,j(e)} d_{i,j(e)}^L \right) \tag{1}$$

where the "employer probability," $p_{i,j}$, is the expectation that person i works at firm, or job, j. For the employer matched sample, which requires at least one match, we only consider jobs with a predicted score surpassing the minimum threshold (see Section 3.2) and use these to calculate probabilities. For the rare case of workers with multiple predicted job matches, we normalize the predicted scores to sum to unity, so that for J jobs, $\sum_{j=1}^{J} p_{i,j} = 1$. [37] Note that weighting is done on a per-job basis for LEHD commutes, using the modal industry at an employer.

The "establishment probability", $p_{i,j(e)}$, is the probability that person i works at establishment e belonging to firm j. For a given job with E_j workplaces, the probabilities are exhaustive, so that $\sum_{e=1}^{E_j} p_{i,j(e)} = 1$. The distance from i's residence to an establishment e (or the indicator of being in the same county) is $d_{i,j(e)}^L$. All candidate establishments with complete distance information are included in the calculations, while those with incomplete information are dropped from calculations.

[36] Unless otherwise noted, these workplaces are based on our own geocoding of the ACS write-in address response. For comparisons meant to align with ACS public use statistics, we also compute an ACS commuting measure that uses the "edited" ACS place of work geography as assigned by ACSO. The results are not substantially different.

[37] For example, suppose that a worker is linked to three employers, A, B, and C with match scores of 0.15, 0.85, and 0.95 respectively, with a match cutoff of 0.8. Employer A fails to make the cutoff and so that job does not contribute to the commuting comparison. Employers B and C both contribute, but with probabilities normalized by their sum, or 0.47 and 0.53.

Table 4: Joint Distribution of ACS and LEHD Employment Status

	Link to LEHD in 3-quarter window	
ACS	Employed	Not employed
Employed	43.1	3.7
Not employed	11.5	41.8

Notes: Sample defined in line 2 of Table 2, representing 667,000 ACS respondents. All numbers are cell percentages. LEHD jobs must have earnings in either the ACS response quarter or the previous or subsequent quarters.

For example, consider a person i matched with two employers, Employer 1 and Employer 2. The analysis sample will contain probabilities associated with each possible employer, $p_{i,j}, j \in 1, 2$. Within each employer j, $p_{i,j(e)}$, give the probability that person i works at establishment e. The expected commute for person i to each employer j, $d_{i,j}$ is computed as $d_{i,j} = \sum_e p_{i,j(e)} d^L_{i,j(e)}$, and the expected commute for person i across all employers is thus $d^L_i = \sum_j p_{i,j} d_{i,j}$.

When multiple establishment matches are possible for a given employer-level match, or $E_j > 1$, we consider several schemes based on different assumptions for calculating establishment assignment probabilities, or $p_{i,j(e)}$. Naïve "uniform" weighting simply assigns equal weight to each establishment, or $p_{i,j(e)} = 1/E_j$. Slightly less naïve, weighting by "establishment size" uses establishment workforce size as weights (but still does not use commute distance). "U2W" weights are computed based on the methods described in Section 2.2, and take into account distance to a worker's residence and size of candidate establishments. At most 10 different establishments receive U2W weights, the remainder being assigned zero weight. We also report results when assigning a unity weight to the establishment with the highest U2W weight, or the "modal" establishment, and zero to all others (randomly allocating ties with an implied minimum weight of 0.1). As a counterpart to using only establishment size, we also report statistics when giving the "closest establishment" unity weight. Finally, leveraging the matching exercise described earlier, we also report statistics that use "match" weights from the establishment matching probabilities. Among those establishments with a sufficiently high score to be deemed a match, the matcher weights are normalized to sum to unity. Note that all but the last of these methods would be feasible to implement in LEHD processing, while the last (requiring links to ACS) could play a role in model development and evaluation.

As noted in Section 3.3, we define the sample for some tables and figures as "distance restricted," meaning that we require that all persons have $d^A_i < 200$mi and $\min\{d^L_{ij(e)}\} < 200$mi, $\forall e \in i$. For LEHD, the restriction compels at least one establishment over all employer matches to have a commute distance of less than 200 miles.

4.2 Joint Distribution of Employment Status

We start in Table 4 by comparing employment status, a prerequisite for commuting, in the overall sample of 667,000 employment eligible persons. Recall that we limit the ACS employment definition to be compatible with LEHD coverage and utilize a three quarter window for LEHD jobs around the quarter including the ACS reference date. We find that 43.1 percent agree on employment, and 41.8 percent agree on non-employment, a total of 84.9 percent. Off the diagonal, we find more instances of persons with LEHD employment but no ACS job (11.5 percent) than the opposite (3.7).[38] The share employed in both files would have been 1.3 percent lower if the LEHD employment window had been limited to only the response quarter.[39]

4.3 Differences Attributable to Person and Job Frames

Table 5 presents the average commute distance and within-county commute rate for different sample restrictions imposed on the ACS-LEHD data. Panels A and B present d^A and d^L, with the statistics in each panel weighted to public use tabulations of ACS and LODES, respectively.[40] Tabulations in Panel A use "edited" geography - the same as the ACS public-use data, while those in Panel B use U2W estaliblishment probabilities and uniform employer probabilities in the case of persons with multiple jobs.

In each panel, we begin with the public use statistics (and a snapshot from the LEHD Infrastructure Files), repeated from Table 1, and narrow the record set to our person-employer matched sample. Following Tables 2 and 3, each bold row definition within a panel is a further refinement of the previous definition and sub-rows provide a breakdown by the completeness of workplace geography.[41]

Panel A of Table 5 shows only small differences in commuting statistics for the ACS microdata compared with the public use statistics, which have a rate of 0.725. The *ACS employed* sample, the subset linked to and LEHD job, regardless of employer match, (in the third bold row), and the *Employer Match* sample have rates of 0.698, 0.693, and 0.694, respectively. The distance restriction reduces average commute distance from 13.9 to 10.1 miles, but has little impact on the

[38] For a comparable analysis using LEHD and the Current Population Survey, though without the use of job level matching, see Abraham et al. [2013]. Looking at the length of LEHD jobs, Abraham et al. [2013] find that workers with short duration jobs in LEHD are especially unlikely to report those jobs on the survey.

[39] Appendix Table A1 provides a version of this table based on LEHD employment only in the response quarter. Our three quarter window for LEHD jobs increases the percentage classified as employed in both frames from 41.8 to 43.1 percent. Alternately, if we define ACS employment to include all those who reported being employed at work, even in other worker classes, the employed agreement rate increases to 46.1 percent.

[40] We calculated weights only for the first row of each microdata sample and then use those weights for subsequent rows. Specifically, for Panel A, we weight "ACS Employed (with PIK)" to the 5% PUMS from ACS for 2007-2011. Likewise, for Panel B, we weight "LEHD Employed (Link to ACS with PIK)" to LODES from 2009. For both the ACS and the LODES, we calculate weights for cells stratified by NAICS sector, age, and sex.

[41] The two sub-definitions break out the two statistics by (i) those respondents who have complete workplace geographic information and (ii) those who are missing some geographic characteristics of their respective place of work. For the edited ACS workplace location, state and county are always available allowing us to always commute a within-county commute rate. Workplace tract is occasionally missing, precluding the calculation of average commute distance. For LEHD, complete geography contains only persons where a distance can be calculated to all potential establishments across all jobs.

Table 5: Commute Analysis by Jobs Sample

	(1) Observations (People)	(2) Average Commute Distance (miles)	(3) Within-County Commute Rate
Panel A. Weighted to Public-Use ACS			
ACS Public-Use			0.725
ACS Employed (with PIK)	311,000	14.6	0.698
Complete geography	248,000	14.6	0.703
Incomplete geography	64,000	NA	0.681
ACS Employed and LEHD Job	287,000	14.6	0.693
Complete geography	230,000	14.6	0.698
Incomplete geography	58,000	NA	0.675
ACS-LEHD Employer Match	226,000	13.9	0.694
Complete geography	187,000	13.9	0.697
Incomplete geography	58,000	NA	0.676
ACS-LEHD Employer Match (distance restricted)	158,000	10.1	0.708
Complete geography	135,000	10.1	0.708
Incomplete geography	23,000	NA	0.703
Panel B. Weighted to Public-Use LODES			
LODES Public-Use			0.549
LEHD			0.553
LEHD Employed (Link to ACS with PIK)	364,000	61.3	0.530
Complete geography	312,000	59.0	0.572
Incomplete geography	52,000	98.9	0.284
ACS Employed and LEHD Job	287,000	52.9	0.539
Complete geography	237,000	50.8	0.574
Incomplete geography	50,000	86.9	0.311
ACS-LEHD Employer Match	226,000	49.1	0.547
Complete geography	199,000	47.3	0.577
Incomplete geography	27,000	80.2	0.327
ACS-LEHD Employer Match (distance restricted)	158,000	30.8	0.608
Complete geography	154,000	30.3	0.614
Incomplete geography	4,000	49.7	0.355

Notes: Panel A uses ACS edited commute distance to calculate average commute distance and within-county commute rate. Panel A weighted to ACS 5-year Public Use Microdata Sample for 2007-2011 and Panel B weighted to LODES for 2009. See Table for definitions of ACS Public-Use, LODES Public-Use, and LEHD. See text and Tables 2 and 3 for definitions of microdata samples as well as the distance restriction. "Complete geography" indicates the subset of person records where commute statistics may be calculated for all workplaces, while "incomplete geography" is the complement. Jobs are aggregated to the person level in Panel B using uniform probablities.

within-county commute rate. For the ACS, we can see that restrictions in the frame do not lead to qualitatively important changes in the statistics of interest.

Panel B of Table 5 shows that changes in the frame produce minimal differences for within-county commute rates in the LEHD, though average commute distances are more sensitive, and certain subgroups have widely differing statistics. The third bold row, which limits the microdata to LEHD jobs linked to the ACS at a person level (matches on PIK, regardless of ACS employment status), has a within-county commute rate of 0.530.[42] The similarity with the LEHD/LODES statistics suggests that people sampled in the ACS have comparable LEHD commutes to those in the universe of jobs in the LEHD, though when ACS responses have incomplete address information, commute distances are substantially higher. Further restricting analysis to the *Employer Match* sample yields a similar within-county commute rate of 0.547. The only significant difference in commute statistics within the LEHD measures can be observed for the *distance-restricted Employer Match* sample, with an increase of the within-county commute rate to 0.608 (and a reduction of commute distance from 49.1 to 30.8 miles). Among out-of-county commutes, LEHD commutes are substantially longer than ACS commutes. We discuss this feature later in the context of establishment non-reporting (see Section 4.6).

The implication from Table 5 is that (1) the longer commutes seen in LODES are also observed in a sample of matched jobs, and (2) the restrictions on a person frame, employment status, and a job frame, do not appear to substantially bias commuting statistics, especially for shorter, within-county commutes. These findings suggest that differences may be due to the remaining design differences for home and workplace assignment, discussed in Section 2.2. The lack of significant differences in the sample frame also suggests that any findings for the matched sample may be broadly applicable to ACS and LEHD commuting statistics. We do highlight the stark differences in commuting distances for those ACS responses with incomplete address information, which we will not be able to address in this study.

For the remainder of this study, we focus on the employer and establishment matched samples restricted to those commuting less than 200 miles (labeled "distance restricted"). We first present statistics for several methods of assigning workplace location. Then, we decompose the remaining differences in commute statistics to disagreement in a range of factors. Lastly, we explain how those features may relate to the design differences.

4.4 Commuting Statistics by Home and Workplace Definitions

Unweighted commuting statistics for the employer and establishment matched samples are reported in Table 6. Columns (1) and (3) are based on the *distance-restricted Employer Match* sample, and Columns (2) and (4) are based on the *distance-restricted Establishment Match* sample. Rows (1) and (2) tabulate ACS statistics, d^A, for "geocoded" and "edited" workplaces, based on the GAL geocoding and the ACSO location assignments respectively. The remaining rows present LEHD-based commuting statistics for several methods of calculating establishment assignment

[42] See Table 4 for the correspondence of ACS and LEHD employment status for this set of persons.

Table 6: Commuting Statistics for Matched Samples (distance restricted)

	(1)	(2)	(3)	(4)
	Within-County Commute Rate		Average Commute Distance (miles)	
	Employer matched sample	Establishment matched sample	Employer matched sample	Establishment matched sample
ACS				
Geocoded	.6975	.7044	9.94	9.84
	(.0012)	(.0015)	(.0374)	(.0407)
Edited	.7102	.7063	11.23	9.93
	(0011)	(.0015)	(0.042)	(0.039)
LEHD				
Uniform	.5558	.5807	35.11	30.68
	(.0012)	(.0015)	(0.234)	(0.269)
Establishment Size	.5768	.6037	32.59	27.41
	(.0012)	(.0015)	(0.329)	(0.352)
Unit-to-Worker (U2W)	.6196	.6461	24.12	18.82
	(.0012)	(.0015)	(0.369)	(0.346)
U2W Modal	.6396	.6699	21.95	16.75
	(.0012)	(.0015)	(0.640)	(0.420)
Closest Establishment	.6957	.7216	12.57	9.65
	(.0012)	(.0015)	(0.069)	(0.082)
Match	NA	.6954	NA	10.89
		(.0015)		(0.086)
Number of Persons (rounded to 1,000s)	158,000	92,000	158,000	92,000

Notes: See Table 3 for sample definitions. All statistics require complete geographic data for the contributing records, which reduces the ACS Edited sample size to 135,000 and 83,000 for the employer and establishment matched samples, respectively. Standard deviations in parenthesis.

probabilities, $p_{i,j(e)}$. For Columns (2) and (4), the set of feasible establishments is restricted to those deemed most likely by the matching process. Each successive assignment methodology (except for the last) results in shorter commutes.

Figure 3 graphs the distribution of d^A and d^L using the *distance-restricted Employer Match* sample with the "Geocoded" workplaces for ACS and the U2W establishment probabilities for LEHD. Both distributions have most of their mass towards zero, with no other mass points and most commutes under 50 miles. LEHD commutes exhibit a longer right tail, with d^L showing more mass for commutes over 20 miles.

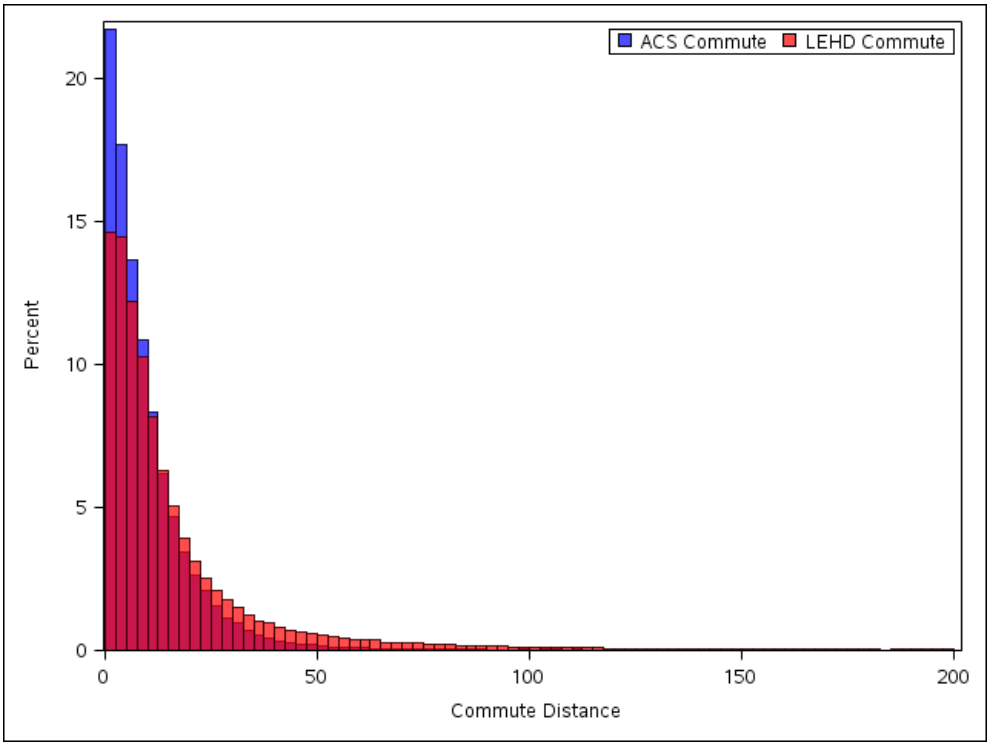


Figure 3: Distribution of Commute Distances in LEHD and ACS

Notes: See Table 3 for definition of *distance-restricted Employer Match* sample, with $N = 158,000$. ACS commutes use "edited" workplaces. LEHD commutes use U2W establishment probabilities for workplaces and, for the case of multiple jobs, use normalized matcher probabilities for employers. The X-axis is capped at 200 miles.

4.5 Decomposition of Remaining Differences by Design Factors

We now consider how design factors in the microdata contribute to the remaining differences. Table 6 gives an indication that the establishment matching and the uncertainty underlying the workplace location in the LEHD data may be contributing to the differences depicted in Table 5. The factors we will consider include the public-use record weighting (LODES or ACS), the matched sample (employer or establishment), the workplace assignment, and the residence assignment. These

are summarized in Table 7. The boldfaced attribute in each column denotes the change from the previous column (from left to right). For the configuration summarized in each column we compute the mean commute distance and the within-county commute rate. Bookending the table, Columns (1) and (8) correspond to the *distance restricted Employer Match* sample used for statistics in Table 5, Panels A (for ACS) and B (for LEHD), respectively. Both configurations weight that sample to the respective public use statistics. Column (1) uses U2W assignments to LEHD candidate establishments and LEHD residences from the CPR, while Column (8) uses ACS edited place of work and the ACS place of residence.

Table 7: Progression of Configurations for Commuting Statistics

Factors	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weighting	LODES	LODES	LODES	LODES	LODES	**ACS**	ACS	ACS
Sample	Employer	**Estab.**	Estab.	Estab.	Estab.	Estab.	Estab.	**Employer**
Workplace	U2W	U2W	**Match**	Match	**ACS**	ACS	**ACS (edit)**	ACS (edit)
Residence	CPR	CPR	CPR	**ACS**	ACS	ACS	ACS	ACS
Obs	158,000	92,000	92,000	92,000	92,000	92,000	83,000	135,000

Notes: Bolded text displays the factor that changes in each configuration. For definitions of *distance-restricted Employer Match* and *distance-restricted Establishment Match* samples, see Table 3. For details, see text.

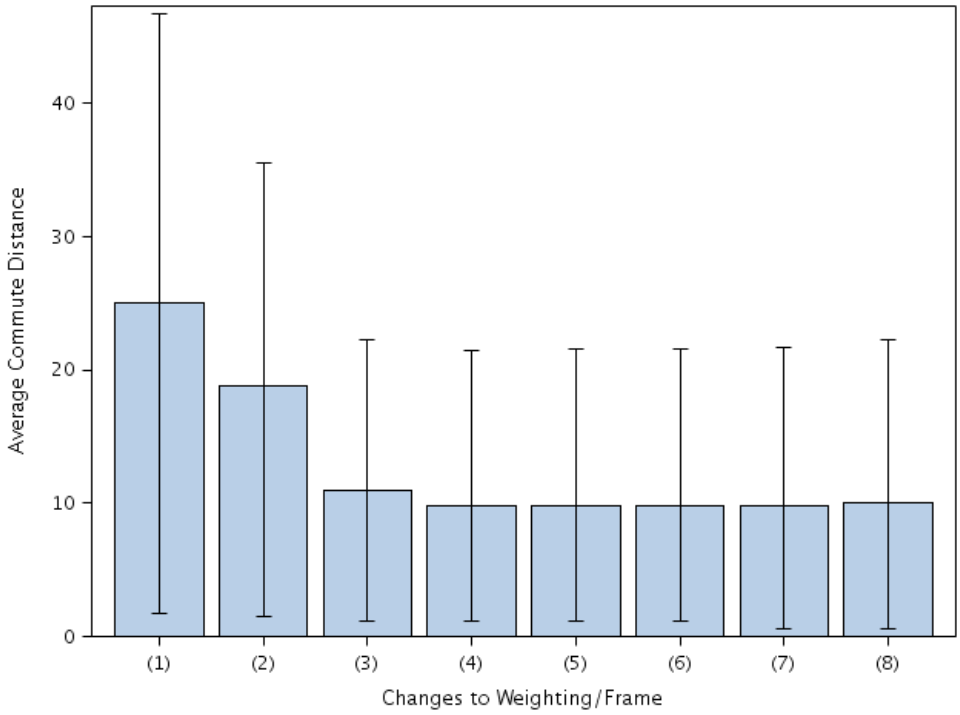
The difference between Columns (1) and (2) highlights the role of workplace address concordance by restricting to the *Establishment Match* sample, which requires that at least one LEHD establishment match to the ACS workplace response. Non-matching LEHD establishments are discarded. Column (3) investigates the role of the linkage model by replacing predicted probabilities based on the U2W framework [Stephens, 2007] with predicted probabilities based on probabilistic matching on name and address information. Column (4) switches place of residence from the LEHD's source on the CPR to the recorded ACS location, allowing us to assess whether the timing and quality of residential address information plays a role. Column (5) replaces probabilistic assignment of LEHD workplace(s) with the deterministic ACS workplace response.[43] Column (6) re-weights records to the ACS employed population, rather than LODES. Column (7) switches workplace location from the geocoded ACS location to the edited ACS location, as it is used for public use statistics.[44]

For each column in Table 7, Figures 4 and 5 depict the average commute distance and the within-county commute rate, respectively.[45]Both figures show that those matching an establishment have

[43]Given that Column (5) uses the *Establishment Match* sample, based on close concordance of name and address of the establishment, the difference between Columns (4) and (5) reflects the "residual" from the matching exercise, translated into a distance measure. Specifically, in cases where one or more establishments at an employer match to an ACS response, but are not exactly located at that address, there will be a difference between the two geocoded locations. As we will see, the contribution from this change is negligible.

[44]The ACS workplaces used in Columns (4)-(6) are for the GAL geocoding of the write-in address. The workplaces used in Column (7) are for the ACS "edited" location, based on ACSO geocoding as well as the resolutions of experienced staff. The staff incorporate outside information to make educated inferences about the place of work if the self-reported address is not valid.

[45]Note that Columns (1) and (8) do not exactly match the corresponding statistics in Table 5 (the last bold rows



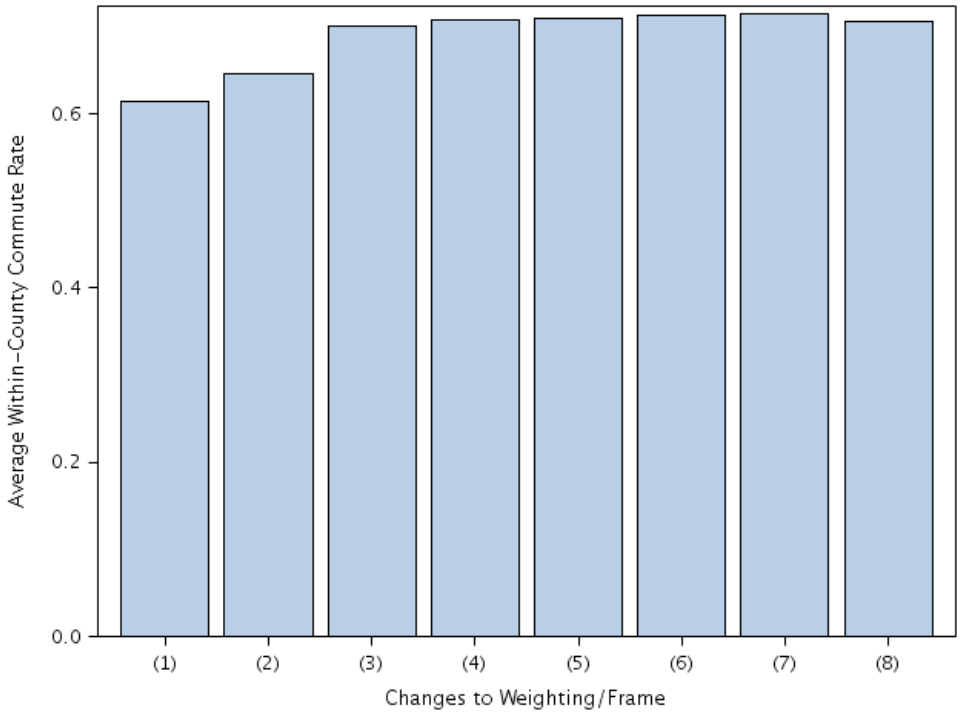
Factors	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weighting	LODES	LODES	LODES	LODES	LODES	**ACS**	ACS	ACS
Sample	Employer	**Estab.**	Estab.	Estab.	Estab.	Estab.	Estab.	**Employer**
Workplace	U2W	U2W	**Match**	Match	**ACS**	ACS	**ACS (edit)**	ACS (edit)
Residence	CPR	CPR	CPR	**ACS**	ACS	ACS	ACS	ACS

Figure 4: Average Commute Distance by Changes in Sample

Notes: Bars denote average commute distances for each sample in miles. Solid lines denote the 10th and 90th percentiles. See Table (7) and the text for a descrition of the different samples. All samples restricted to observations where at least one LEHD job has at least one estbalishment less than 200 miles from the LEHD residence. Sample sizes for each stage are as follows: column (1) 158,000, columns (2)-(6) 92,000, column (7) 83,000, and column (8) 135,000.

shorter commutes (bringing them more in line with ACS). With at least one establishment match, the average commute distance falls from 25.0 to 18.8 miles and the within-county commute rate rises from 0.61 to 0.65 (almost half of the remaining difference). Commutes shorten again when moving from U2W to matcher probabilities in Column (3), reducing average commutes to 10.9 miles and increasing the within-county commute rate to 0.70. Switching the source of residential addresses (Column 4) further reduces commute distance to 9.8 miles and raises within-county commute rate to 0.71. The remaining factors make little difference.

in each panel) because of differences in the weighting, with distance being more sensitive. While Table 5 weights records in the first "employed" sample rows to the public-use data, and uses those weights for subsequent statistics, the configurations described in Table 7 weight records from the *distance restricted Employer Match* sample to the public-use data.

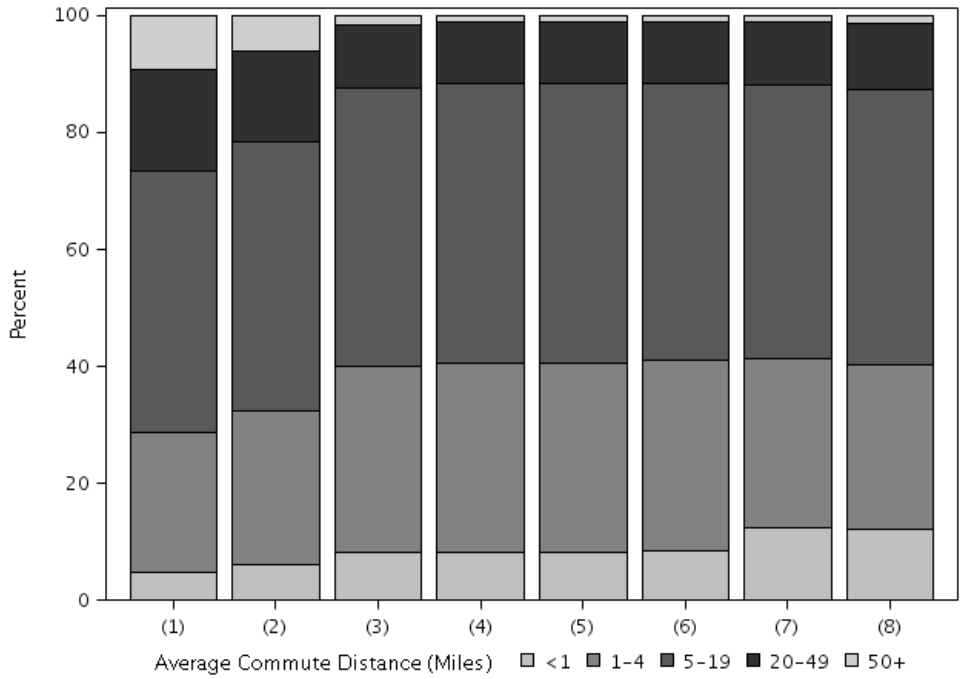


Factors	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weighting	LODES	LODES	LODES	LODES	LODES	**ACS**	ACS	ACS
Sample	Employer	**Estab.**	Estab.	Estab.	Estab.	Estab.	Estab.	**Employer**
Workplace	U2W	U2W	**Match**	Match	**ACS**	ACS	**ACS (edit)**	ACS (edit)
Residence	CPR	CPR	CPR	**ACS**	ACS	ACS	ACS	ACS

Figure 5: Within-County Commute Rate by Changes in Sample

Notes: Bars denote share of sample commuting to a workplace within the county of residence. See Table (7) and the text for a descrition of the different samples. For the underlying values, see Appendix Table A5.

Figure 6 provides information on the distribution of commute distances that are driving the changes in the average commute distance. Each bar depicts the share of commutes of less than one mile, one to four miles, 5-19 miles, 20-49 miles, and greater than 50 miles. Decreases in the average commute distance as one moves from left to right are accompanied by shifts in the distribution of commutes from longer to shorter. Specifically, commutes of greater than 50 miles decrease drastically, as previously hinted at by Figure 3. The increase in the fraction of workers making the shortest commutes once the ACS edits are applied in Column (7) is suggestive of a specific feature of the place of work edits. ACS respondents who say that they work from home (inclusive of "teleworking") may still provide an employer name and address. Whereas in Column (6) the commute distance is based on the geocoding of reported employer address, the ACS edit used in Column (7) applies an edit rule setting the workplace to the home location (giving a commute distance of zero). This edit mainly affects commutes that were already relatively short, as can be



Factors	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weighting	LODES	LODES	LODES	LODES	LODES	**ACS**	ACS	ACS
Sample	Employer	**Estab.**	Estab.	Estab.	Estab.	Estab.	Estab.	**Employer**
Workplace	U2W	U2W	**Match**	Match	**ACS**	ACS	**ACS (edit)**	ACS (edit)
Residence	CPR	CPR	CPR	**ACS**	ACS	ACS	ACS	ACS

Figure 6: Distribution of Average Commute Distance by Changes in Sample

Notes: Bands denote share of each sample within each commuting distance bin. See Table (7) and the text for a descrition of the different samples. For the underlying values, see Appendix Table A5.

seen in Figure 6.

We have shown that three key differences in the design of the ACS and LEHD commute statistics can account for nearly half of the difference in commute statistics: differences in the worksite frame, the inherent uncertainty in establishment assignment when exact work location is unknown, and differences in residence location. The next sections consider each of these factors in turn, attempting a deeper understanding of how and why each contributes.

4.6 How Differences in the Worksite Frame Contribute to Longer LODES Commutes

When restricting the sample to only those ACS workplace responses that match to an establishment in the LEHD, the average commute distance falls from 25.0 to 18.8 miles (Table 7 and Figure 4). The drop in distance implies that ACS jobs with no matched LEHD establishments have longer

commutes. Failure to match at the establishment level could occur for a number of reasons, including: inaccurate or incomplete address fields, different perspectives on workplace definition, and incomplete establishment reporting. We briefly discuss two of these before focusing on the last one.

Incomplete geography. The last main row of Panel B in Table 5 disaggregates commuting statistics for the *distance-restricted Employer Match* sample by whether geographic information is complete for all LEHD candidate establishments for a given employer. If any of an employer's establishments have incomplete geographic information, all establishments are classified as "incomplete," and any statistics are computed only over those establishments within this group that have complete information. For the 4,000 records with some incomplete geographic information (out of 158,000), average commutes are 49.7 miles, relative to 30.3 for the complement and 30.8 for the entire sample. It is not immediately clear why they would have longer commutes. One possibility is that the subset of establishments for which employers *do* in fact report accurate locations systematically omits locations that are closer to residential areas (for example, rural routes, in areas where travel distances are longer, can be more challenging to geocode). However, they constitute a small share of records and contribute little to overall ACS/LODES differences.

Workplace definitions. Second, differences in workplace definition are likely for certain jobs and industries. The establishment match rate for ACS respondents who say they work at home is substantially lower than that of other ACS respondents - 10.56 versus 36.74 percent, but these records account for only 2.3 percent of the total. Another case where establishment definitions in LEHD may not conform with the notion of place of work for a survey respondent is for industries with widely dispersed employment. The industries with the highest commuting distances for LEHD in the employer matched sample, as reported in Appendix Table A4, have especially low establishment match rates (see Appendix Table A3). These industries, including mining, transportation and warehousing, and administrative support and waste management, tend to have a mobile, wide ranging workforce.

Establishment non-reporting. Third, we present evidence on how establishment non-reporting in LEHD may contribute to longer commutes. Along with information on the primary workplace, employers subject to QCEW reporting are asked to submit a Multiple Worksite Report (MWR, see Section 2.2) if workers are engaged in multiple economic activities (industries) or have multiple worksites (if secondary worksites account for 10 or more employees). As of 2017, 25 states and the District of Columbia mandate a response, the other 25 states have voluntary reporting. Non-compliance is estimated at 5.61 percent [Spear, 2011]. States where compliance is mandatory have a non-compliance rate of only 3.66 percent, while it rises to 7.90 percent in voluntary states. Non-compliance is especially high for local governments, at 8.94 percent.

Patterns of non-reporting are varied. Some employers fail either to report new worksites for a given period of time or to promptly notify the state when a worksite is no longer in operation. Some large employers simply fail to report multiple worksites altogether. Experience with quality analysis of LEHD has found cases of large employers (i.e. an SEIN with many workers), where employment would be expected to be distributed across many locations (e.g. school districts, home

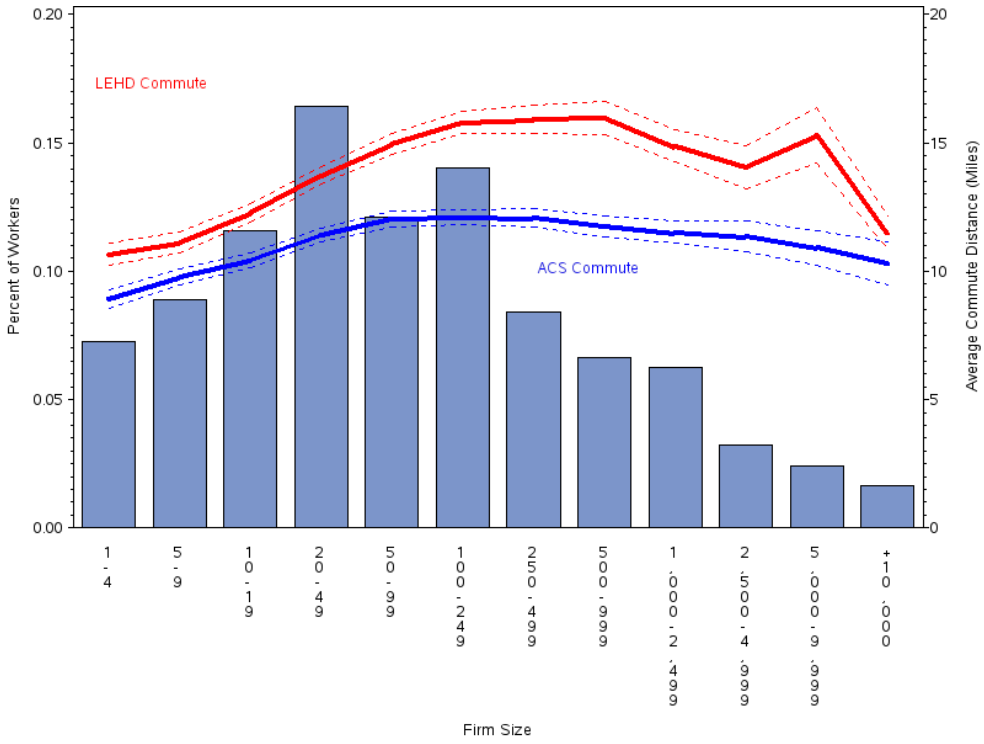


Figure 7: Commute Distance for Single-Unit Employers by State Firm Size

Notes: $N = 94,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across state firm sizes corresponding to the left vertical axis. Solid lines denote average commute distance by state firm size corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is a subset of the *distance restricted Employer Match* sample, defined in 3. For this subset, ACS workers match to only one LEHD employer and the employer has only one establishment. ACS commutes use geocoded workplaces and LEHD commutes use U2W establishment probabilities.

care providers), attribute all employment to the employer's headquarters or account office. Establishment match rates for state and local governments are 30.7 and 32.9 percent, respectively, compared to 37.1 percent for private sector establishments (Appendix Table A2). For large states, unreported worksites will contribute to longer LEHD commutes because the single, reported location will be further from workers' homes, on average, than the places they actually perform their duties.

We investigate this issue by focusing on ACS respondents with only one candidate establishment in the LEHD (and only one employer). The LEHD program has no direct information on whether an employer failed to report multiple-worksites when it should have (although longitudinal changes in reporting are identifiable). However, non-reporting is more likely in the case of a single-unit employers with a large workforce. While small, single-unit employers are routine, one would expect large employers to be more likely to have multiple establishments. If the incidence of non-reporting among single-unit SEINs is correlated with firm size, and non-reporting induces longer average

commutes, then commute distance should be positively correlated with firm size.

Figure 7 plots the average commute distance for LEHD (d^L) and ACS (d^A), respectively. While distance increases for firm sizes below 100 in LEHD as well as ACS as well, and $\Delta d = d^L - d^A$ is positive for all bins, the gap Δd increases with firm size: A gap of about two miles for firms with one to nine workers grows to a gap of four miles for firms over 100 workers. Over half of LEHD jobs are at single-unit employers and almost half of these are at larger firms (over 100 workers), suggesting that non-reporting (along with the other mechanics described above) is a significant contributor to differences in worksite frame and commute distance.

However, non-reporting should also reduce the number of candidate establishments available to be matched to the ACS response. At the limit, complete non-reporting by an employer should reduce the available number of establishments, conditional on a match on employer name, to a single establishment. This is not the case in the *Employer Match* sample. The number of candidate establishments is actually slightly higher in the non-matched sample (46 vs. 43).

4.7 How Uncertainty in Establishment Assignments Contributes to Longer LEHD Commutes

Replacing the imputed work location from the U2W with the matched establishment based on name and address match (Column 3 of Table 7) allows us to investigate the role of uncertainty in workplace location, inherent to the LEHD infrastructure. By construction, the analysis is limited to the *Establishment Match* sample. In Figure 8, we plot sample shares and commute distances d^L and d^A against the number of establishments per employer. Approximately 60 percent of workers are at an employer with only one establishment, while a small share have over 200 establishments. The single-unit share in this sample is moderately larger than in LEHD overall (about 56 percent).

First, note that d^A is relatively constant across bins, at around 11 miles, regardless of the count of employer establishments. For single establishment employers, d^L is almost exactly equal to d^A. This correspondence is not quite true by construction – establishment matches may not be exactly co-located and worker's residences may still differ between ACS and LEHD. For multi-unit employers with only two to five establishments, d^L rises to about 30 miles - a 20 mile gap from ACS. The LEHD commute distance continues to rise with establishment count up to about 40 miles for those with over 200 units. We obtain similar results using the *Employer Match* sample (see Appendix Figure A8), but with a larger gap for single establishment employers (explained in Section 4.6 above) and a higher baseline and steeper slope for d^L at multi-unit employers.

Table 6 helps explain the differences shown here and suggests some further analysis. Commuting statistics for d^L based on the U2W assignment probabilities were closer to d^A than those based on the "uniform" and "establishment size" probabilities. However, when down-weighting lower probability establishment links (the "U2W modal" and "closest establishment" models), Δd was even closer to zero. Longer commute distances in the *Establishment Match* sample suggests that the U2W impute probabilities for distant establishments may be biased upwards, indicating a potential misspecification of the model or inappropriate constraints for selecting candidate establishments

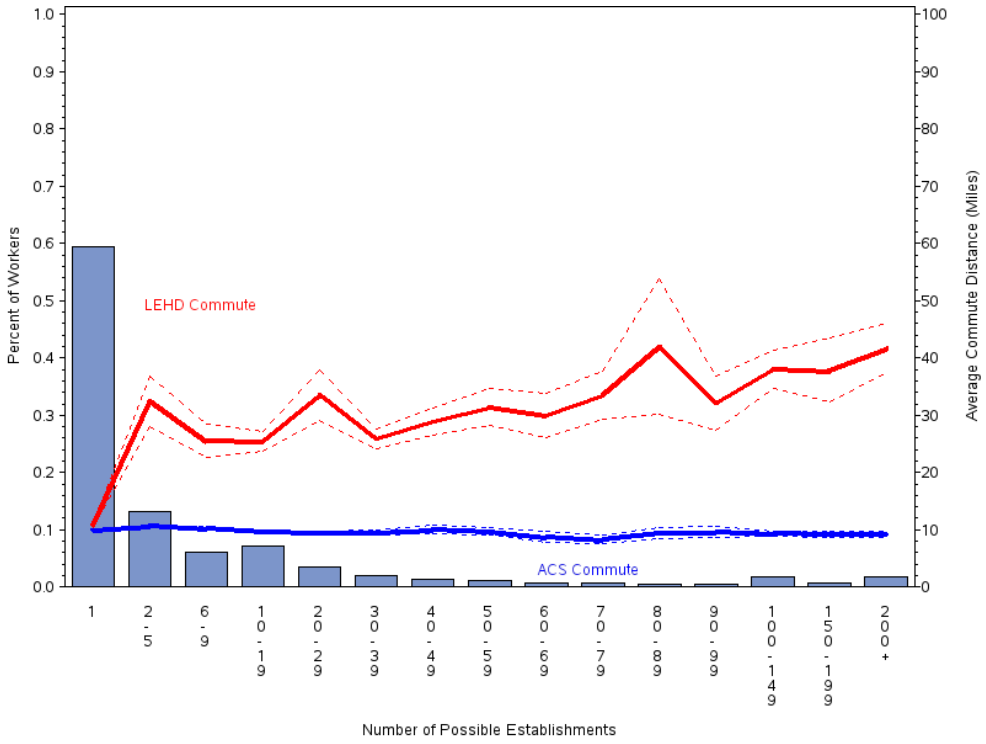


Figure 8: Commute Distance by Number of Establishment Candidates for Persons who Match to at Least One Establishment with LEHD Commutes Using U2W Establishment Probability

Notes: $N = 92,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across number of candidate establishments corresponding to the left vertical axis. Solid lines denote average commute distance by possible establishment matches corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is the *distance restricted Establishment Match* sample, defined in 3. ACS commutes use geocoded workplaces and LEHD commutes use U2W establishment probabilities.

for a job.[46]

Figure 9 shows d^L by establishment count classes when computed with "U2W modal" probabilities, i.e., only the establishment with the highest U2W probability receives positive weight. Compared to Figure 8, d^L decreases for most establishment count classes, except for the highest size class, with the strongest reduction affecting firms with a small to medium number of establishments. [47]

These results suggest that a re-evaluation of the linkage model between workers and workplaces might be useful in reducing the discrepancy between ACS and LEHD commute distances. Whereas

[46]The U2W imputation is constrained to only include establishments that exist for the entire duration of a worker's spell at an employer. U2W draws are applied for the entire job spell, implying that there are no transfers between establishments [Stephens, 2007].

[47]We also tested a hypothesis that the reliance of the U2W model on only ten draws, or implicates, might be a poor representation of the expected probabilities. Using multiple LEHD production vintages to increase the number of draws available for a given job, we calculated commuting statistics using these smoother probability distributions. Results showed little difference in average commute distance compared to using the standard 10 implicates.

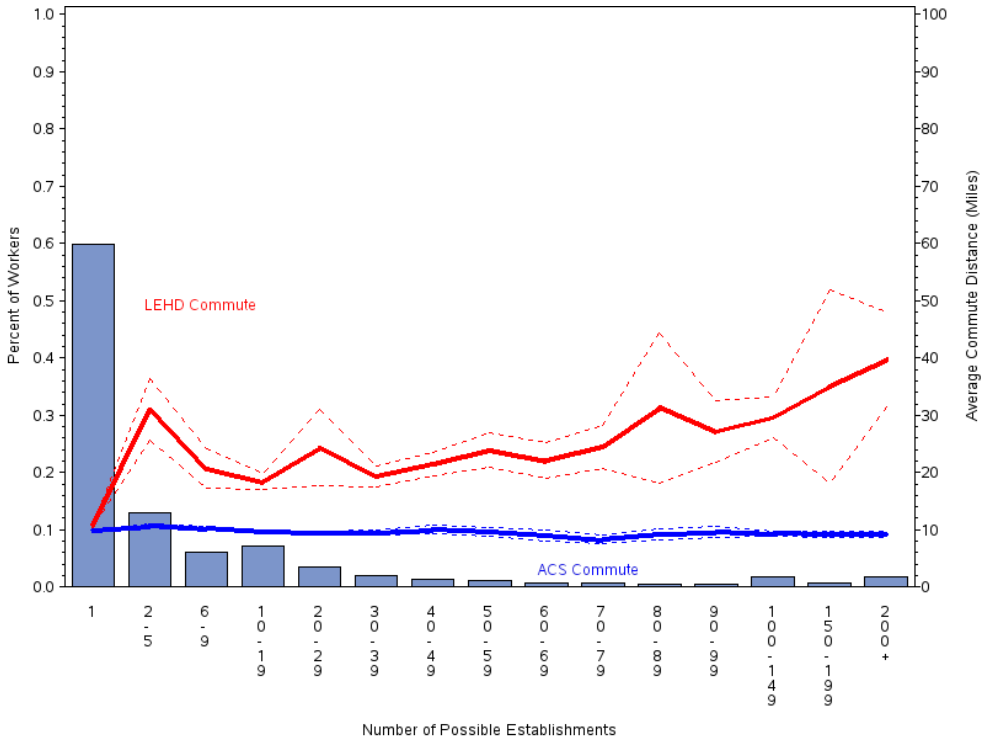


Figure 9: Commute Distance by Number of Establishment Candidates for Persons who Match to at Least One Establishment with LEHD Commutes Using Modal U2W Establishment Probability

Notes: $N = 92,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across number of candidate establishments corresponding to the left vertical axis. Solid lines denote average commute distance by possible establishment matches corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Jobs weighted with matcher weights using logit model. All observations in the sample have at least one job with one establishment less than 200 miles from LEHD residence.

the original model was trained on the universe of workers and employer-reported workplaces in Minnesota (conditional on MWR-related measurement error), the matched ACS-LEHD dataset allows for estimation using a larger, and nationally representative sample. To demonstrate the potential for such a model to replicate the commute distances in ACS, we compute d^L with the matcher probabilities, reported in Figure 10. The results suggest that when the establishment universes of LEHD and ACS overlap, choosing the "right" establishment can explain most of the difference between d^L and d^A.

4.8 How Differences in Residence Location Contribute to Longer LEHD Commutes

Figures 4 and 5 showed a small change in commute statistics when switching from LEHD place of residence (Column 3) to ACS place of residence (Column 4). We provide an explanation here of how

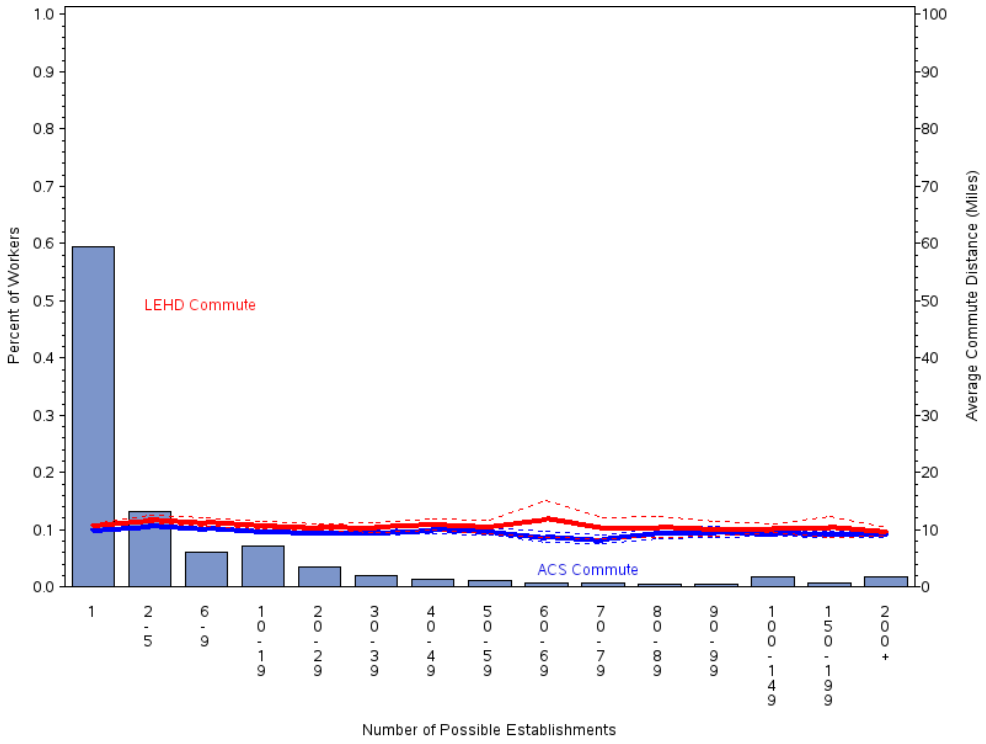


Figure 10: Commute Distance by Number of Establishment Candidates for Persons who Match to at Least One Establishment with LEHD Commutes Using Matcher Probability

Notes: $N = 92,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across number of candidate establishments corresponding to the left vertical axis. Solid lines denote average commute distance by possible establishment matches corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is the *distance restricted Establishment Match* sample, defined in 3. ACS commutes use geocoded workplaces and LEHD commutes use matcher probabilities for establishments.

conflicting residence locations might affect commute distance, though further investigation may be warranted. Figure 11 gives the average commute distances for the *distance restricted Establishment Match* sample, by distance between the ACS and LEHD residences, overlaid onto a histogram of the distance between residential locations.

Over 90 percent of the observations have an exact or near match ($<$ 1 mile) on residence location. The remaining mass in the distribution is almost entirely within 15 miles, indicating that the residence information between the two data sources mostly agrees and that most moves in a short timeframe are local.[48] However, the average commute distances for the ACS and LEHD are increasing as the discord between residences grows.

Why might LEHD commutes rise faster as residences disagree more? While ACS questionnaires collect information from a person's current place of residence, there may be a time gap between when

[48]The Current Population Survey Annual Social and Economic Supplement for 2016 reports a domestic migration rate of 10.7 percent, with only 3.9 percentage points being moves from one county to another.

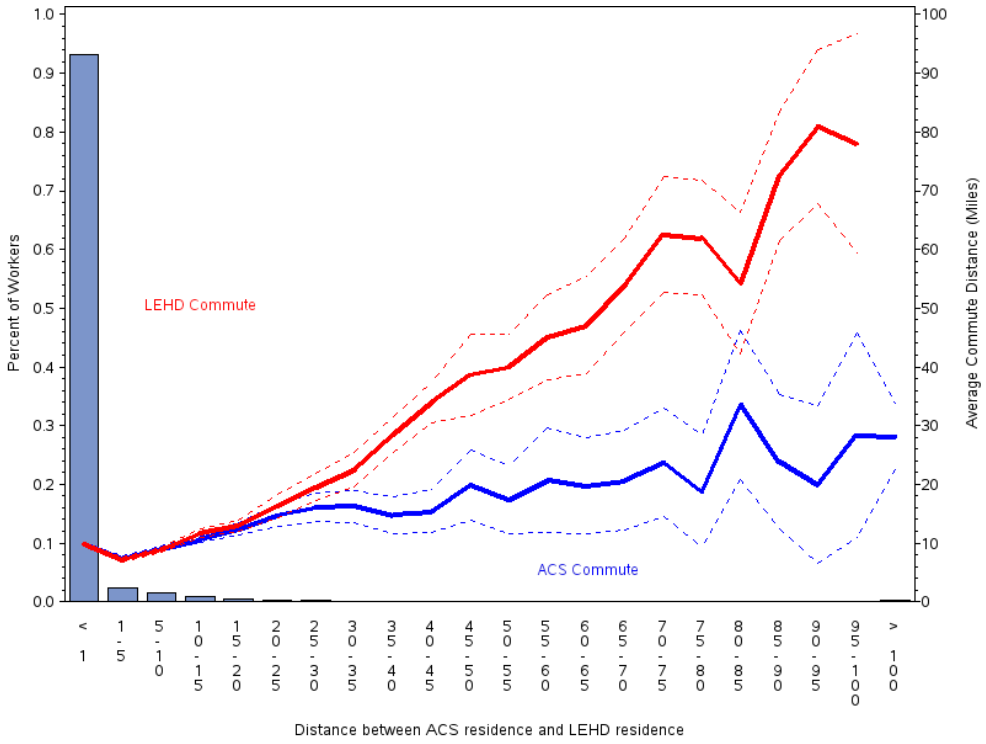


Figure 11: Average Commute Distance by Disagreement of LEHD and ACS Residences for Persons who Match to at Least One Establishment

Notes: $N = 92,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across distances between residences corresponding to the left verticle axis. Solid lines denote average commute distance by distance between residences corresponding to the right verticle axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is the *distance restricted Establishment Match* sample, defined in 3. ACS commutes use geocoded workplaces and LEHD commutes use matcher probabilities for establishments.

LEHD quarterly earnings are measured and when administrative residence data is reported. With LEHD residence recorded only once a year per person, there is the possibility for a long distance move to have occurred and for job and residence information in the same year to be misaligned. While these differences apply to only a small share of jobs, the average distances are large and contribute to the longer LEHD commutes overall. Graham et al. [2016] find that disagreement between the residences of ACS respondents and linked administrative records are more common for younger persons, which is consistent with higher mobility for that group.

5 Discussion

Figure 12, which is an expansion of Figure 1, summarizes our approach to decomposing differences across the two public use datasets and presents the within-county commute rate at each intermediate step. We focus on the within-county commute rate here because it is directly computable from ACS

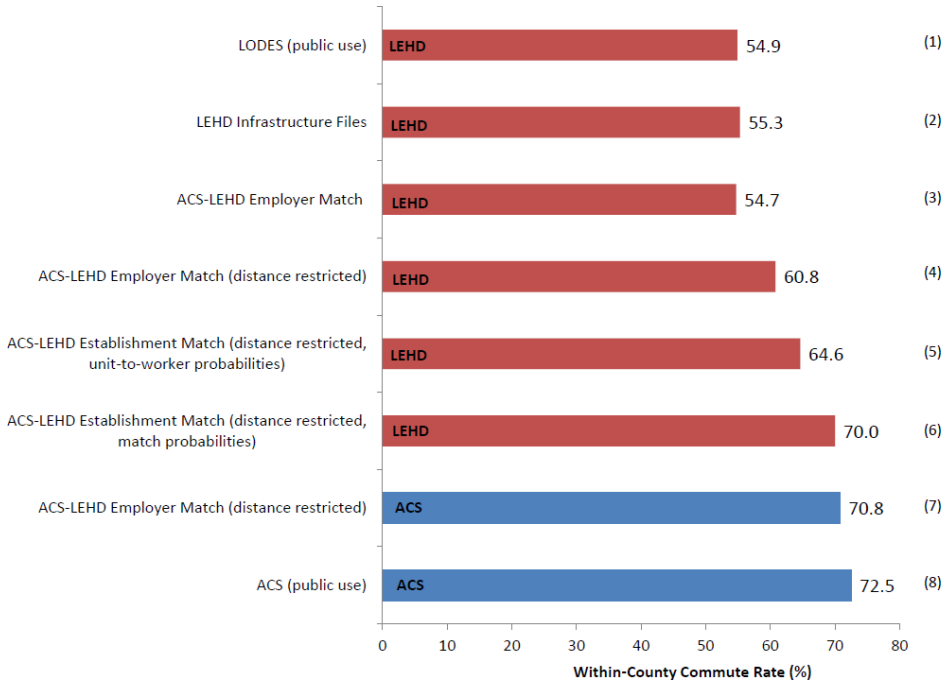


Figure 12: Summary of Within-County Commute Rate by Jobs Sample

Notes: Bars denote within-county commute rate for each sample, with red bars for LEHD home and workplace locations and blue bars for ACS home and workplace locations. From top to bottom, the within-county commute rate for each bar can be found in the following tables: Table 1, Table 1, Table 5, Table 5, Table A5, Table A5, Table 5, Table 1.

and LODES public use data. Within-county commute rates in red use both residence and workplace locations from the LEHD, and those in blue, from the ACS. All statistics are weighted to match the job composition of the respective public use files.

The top and bottom bars, for LODES and ACS, respectively, show a lower within-county commuting rate of 54.9 percent, in LODES (row 1), compared with 72.5 percent in ACS (row 8), a gap of 17.6 points. Reading the top three rows, we find negligible differences in the rate due to changes in the sample of LEHD jobs. Using the LEHD Infrastructure Files microdata (row 2) that serve as an input to LODES, we find little difference from the public use data. Interestingly, restricting LEHD jobs to the ACS-LEHD *Employer Match* sample (row 3) finds a very similar rate of 54.7 percent (though Table 5 showed a modest drop in commute distance).

The major differences are evident in three adjustments to the *Employer Match* sample. First, restricting the *Employer Match* sample to records with ACS and LEHD commutes of less than 200 miles increases the within-county rate to 60.8 percent (row 4), accounting for a third of the gap. Second, further restricting to the *distance restricted Establishment Match* sample yields a local commute rate of 64.6 percent (row 5), accounting for another third of the gap. Third, applying

establishment probabilities to LEHD that correspond to the matcher predictions for ACS workplace response mostly closes the remaining gap to 70.0 percent (row 6). Commute statistics based exclusively on ACS data (in blue) vary little across microdata and public use samples, as is evident from the similarity of rows (7) and (8).

In summary, roughly two thirds of the gap may be mostly attributable to differences in workplace frames (either due to extreme distances or location mismatch), while the remaining third may be due to uncertainty in the assignment of LEHD establishments to workers.

6 Conclusion

In the mid-2000s, two new sources of commuting statistics emerged after the discontinuation of the Census long-form. Both the ACS and LODES, which provide commuting statistics based on survey and administrative data, respectively, have active user bases. However, the two data sources provide different answers with regard to typical metrics, such as commute distance or within-county commute rate. Relative to the ACS, commutes in LODES are distributed over wider distances.

This study has matched ACS respondents' records to their jobs in the LEHD microdata underlying LODES. The ACS is a nationally representative survey, with workers reporting their employer's name and their workplace address. The LEHD microdata relies on the near-universe of employer-provided reports of jobs held by workers, and a separate report of potential worksites for each employer. Personal identifying information, job timing, and employer descriptors facilitate job level matching, which succeeds in almost three quarters of cases. We investigate the sources of the divergence in commuting patterns by looking at sampling, frame discrepancy, and certain measurement issues.

Our results suggest two broad classes of issues contributing to the difference in commute distances: disagreement in workplace locations as reported in survey and administrative data and the missing data problem associated with linking workers to their unique work locations in the administrative data.

Firms may underreport establishments and may also misreport the location of establishments, just as workers may also misreport a workplace, leading, in either case, to a discrepancy in workplace and commute distance. The current LEHD model for assignment of work locations to jobs forces every worker to have a workplace location among the observed LEHD establishments. Underreporting of units will lead to a higher average commute distance, despite the LEHD model favoring shorter commutes. A mere discrepancy between ACS and LEHD locations can have an ambiguous effect on commute distances. Both under- and misreporting will lead a failure to match when using address information. Our findings show that when no LEHD workplace can be matched to a reported ACS workplace, the commute distance, whether measured in the ACS or in LEHD, is higher. When a match is possible, using all available contemporaneous name and address information, the use of establishments selected by the matcher model instead of the current LEHD model reduces the discrepancy substantially. These findings are consistent with underreporting. On the

other hand, firms in the *Establishment Match* sample report as many establishments as those in the complement, for which no establishment matches to the ACS response. This similarity suggests that it is not simply the *number* of establishments reported by firms in the LEHD that is at issue.

Several caveats apply to our comparisons of linked records, which omit unlinked records that may differ in some respects. First, a substantial mismatch remains between ACS reported work status and LEHD observed work status. About 8 percent of ACS employed have no corresponding job, and thus no workplace, in LEHD; about 21 percent of LEHD employed in the matched sample report no work in ACS. In the LEHD, these workers have much higher observed commute distances. Second, among those that agree on work status in both sources, 21 percent have either insufficient information for a match or cannot be matched. Again, the unmatched have much higher commute distances. Match rates vary systematically by industry (Appendix Table A3), for reasons that seem intuitive, but will need to be investigated more closely. Finally, among those that match to an employer, nearly half do not match on address to an employer-reported establishment and again, have longer commutes. These composition differences give some indication that LEHD data, by virtue of its administrative source and its near-universal coverage, may capture more long-distance "commutes," or what are currently interpreted as commutes. Further investigation may be warranted.

The administrative data available to LEHD only rarely provides direct information as to the workplace location of specific workers. For employers that are observed to have more than one establishment within a state, LEHD must represent the likelihood of worker being assigned to various workplaces. Contemporaneous ACS responses could be matched to an LEHD establishment for about 37% of workers in our sample. For these same workers, the LEHD assignments from the Unit-to-Worker (U2W) imputation model resulted in substantially longer commutes. These results suggest that there is scope to improve assignments, but the key data used in this study (ACS address information from an alternate source) are not directly available for most people found in the LEHD. An adjustment or a replacement for the currently used U2W model may need to be developed. Our study also considered possible discrepancies in residence location, and found this to be only a minor contributor to the discrepancy in commute distances.

The set of linked ACS and LEHD jobs produced in the matching exercise and the findings of the study will support further investigation into differences in the public use data and may contribute to quality improvements in both datasets. LEHD data could be used to enhance imputation models for ACS when workplace information is missing or incomplete. For instance, the ACS imputation model for workplace information uses a hot deck model, with donors provided by neighbors of the respondent. One could investigate the value added by expanding the donor pool to similar persons matched from administrative records. The linked ACS and LEHD jobs may also be used to improve LEHD imputation models for allocating workers to establishments at multi-unit employers. Specifically, the *Establishment Match* file could serve as a truth set for re-estimating the U2W imputation model and evaluating its quality. Beyond the set of reported LEHD establishments, the matched set could inform a model to identify cases where reported establishments likely do

not represent that actual distribution of workplaces for an employer. Such cases could be treated differently for internal processing and statistical reporting. The LEHD program may also use the linked file to enhance the model for discriminating between residence locations or to identify residences that seem unlikely in the context of a particular job.

The present study has used microdata linkages to investigate discrepancies in commuting statistics in two widely used datasets, attempting to reconcile the responses provided by workers on their daily commute and those implicitly provided by employers on how far away their workers live. We have identified several promising avenues of investigation and even possible implementation, for a significant part of the data. For other parts, questions remain as to differences in labor force status, timing, and composition.

References

AASHTO. Commuting in America 2013 - Brief 15: Commuting flow patterns. Technical report, American Association of State Highway and Transportation Officials, 2015. URL http://traveltrends.transportation.org/Documents/B15_Commuting%20Flow%20Patterns_CA15-4_web.pdf.

J. M. Abowd, B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock. The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In *Producer Dynamics: New Evidence from Micro Data*, NBER Chapters, pages 149–230. National Bureau of Economic Research, Inc, December 2009. URL https://ideas.repec.org/h/nbr/nberch/0485.html.

K. Abraham, J. Haltiwanger, K. Sandusky, and J. Spletzer. Exploring differences in employment between household and establishment data. *Journal of Labor Economics*, pages 129–172, 2013. URL http://www.nber.org/papers/w14805.

F. Andersson, J. C. Haltiwanger, M. J. Kutzbach, H. O. Pollakowski, and D. H. Weinberg. Job displacement and the duration of joblessness: The role of spatial mismatch. Working Papers No. 20066, NBER, 2014. URL http://www.nber.org/papers/w20066.

T. Evans, L. Zayatz, and J. Slanta. Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics*, pages 537–551, 1998. URL http://www.jos.nu/Articles/abstract.asp?article=144537.

Federal Highway Administration. 2009 National Household Travel Survey user's guide (version 2). Technical report, U.S. Department of Transportation, 2011a. URL http://nhts.ornl.gov/2009/pub/UsersGuideV2.pdf.

Federal Highway Administration. Summary of travel trends: 2009 National Household Travel Survey. Technical Report FHWA-PL-ll-022, U.S. Department of Transportation, 2011b. URL http://nhts.ornl.gov/2009/pub/stt.pdf.

S. Fu and S. L. Ross. Wage premia in employment clusters: How important is worker heterogeneity? *Journal of Labor Economics*, pages 271–304, 2013. doi: 10.1086/668615. URL http://doi.org/10.1086/668615.

G. Gathright, A. Green, M. Kutzbach, K. McCue, H. Monti, A. Rodgers, L. Vilhuber, N. Wasi, and C. Wignall. Employer list linking: Methods, tools, implementation, and robustness. Unpublished manuscript, U.S. Census Bureau, 2016.

M. Graham and P. Ong. Social, economic, spatial, and commuting patterns of dual jobholders. Longitudinal Employer-Household Dynamics Technical Paper No. TP-2007-01, Center for Economic Studies, U.S. Census Bureau, 2007. URL https://www2.census.gov/ces/tp/tp-2007-01.pdf.

M. R. Graham, M. J. Kutzbach, and B. McKenzie. Design comparison of LODES and ACS commuting data products. Working Papers 14-38, Center for Economic Studies, U.S. Census Bureau, Oct. 2014. URL https://ideas.repec.org/p/cen/wpaper/14-38.html.

M. R. Graham, M. J. Kutzbach, and D. H. Sandler. Developing a residence candidate file for use with employer-employee matched data. *2015 Federal Committee on Statistical Methodology Research Conference Proceedings*, 2016. URL https://fcsm.sites.usa.gov/files/2016/03/H1_Graham_2015FCSM.pdf.

M. W. Horner and D. Schleith. Analyzing temporal changes in land-use-transportation relationships: A LEHD-based approach. *Applied Geography*, pages 491–498, 2012. doi: 10.1016/j.apgeog.2012.09.006. URL http://doi.org/10.1016/j.apgeog.2012.09.006.

H. Hyatt. Co-working couples and the similar jobs of dual-earner households. Working Papers 15-23, Center for Economic Studies, U.S. Census Bureau, 2015. URL ftp://ftp2.census.gov/ces/wp/2015/CES-WP-15-23.pdf.

E. Isenberg, L. C. Landivar, and E. Mezey. A comparison of person-reported industry to employer-reported industry in survey and administrative data. Working Paper Number 2013-24, Social, Economic, and Housing Statistics Division, 2013. URL https://www.census.gov/people/io/files/Isenberg%20Landivar%20Mezey%20Industry%20working%20paper%20SEHSD.pdf.

J. Lane, M. Roemer, W. Mix, G. Putnam, W. Almousa, M. OConnell, and G. Foster. An evaluation of the use of LEHD data for transportation planning. Longitudinal Employer-Household Dynamics Technical Paper No. TP-2003-11, Center for Economic Studies, U.S. Census Bureau, 2003. URL https://www2.census.gov/ces/tp/tp-2003-11.pdf.

A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. *International Conference on Data Engineering (ICDE)*, pages 277–286, 2008. doi: 10.1109/ICDE.2008.4497436. URL http://doi.org/10.1109/ICDE.2008.4497436.

B. McKenzie. Who drives to work? Commuting by automobile in the United States: 2013. American Community Survey Reports ACS-32, U.S. Census Bureau, Aug. 2015. URL https://www.census.gov/hhes/commuting/files/2014/acs-32.pdf.

E. Murakami. Understanding LEHD and synthetic home to work flows in "ON THE MAP". Federal highway administration, fhwa office of planning, U.S. Department of Transportation, 2007. URL https://www.fhwa.dot.gov/planning/census_issues/lehd/lehdonthemap.cfm.

NCHRP. A guidebook for using American Community Survey data for transportation planning. Report 588, National Cooperative Highway Research Program (NCHRP), 2007.

Office of Management and Budget. Revised delineations of metropolitan statistical areas, micropolitan statistical areas, and combined statistical areas, and guidance on uses of the delineations of these areas. OMB Bulletin No. 13-01, Executive Office of the President, 2013. URL https://obamawhitehouse.archives.gov/sites/default/files/omb/bulletins/2013/b13-01.pdf.

B. D. Spear. NCHRP improving employment data for transportation planning. Project 08-36, National Cooperative Highway Research Program (NCHRP), 2011.

B. Stephens. *Essays on firm compensation policy and confidentiality protection and imputation in the Quarterly Workforce Indicators*. Ph.d., University of Maryland, College Park, 2007. URL http://gradworks.umi.com/32/41/3241441.html.

D. W. Stevens. Employment that is not covered by state unemployment insurance laws. LEHD Technical Paper No. TP-2007-04, U.S. Census Bureau, 2007.

C. M. Tolbert and M. Sizer. US commuting zones and labor market areas: A 1990 update. Staff Report ERS-AGES-9614WTS, Economic Research Service, 1996. URL http://trid.trb.org/view.aspx?id=471923.

U.S. Census Bureau. County to county commuting flows for the United States and Puerto Rico: 2009-2013. [excel file], U.S. Census Bureau [distributor], 2015. URL https://www.census.gov/hhes/commuting/files/2013/Table%201%20County%20to%20County%20Commuting%20Flows-%20ACS%202009-2013.xlsx.

U.S. Census Bureau. Design and methodology: American community survey. Technical report, United States Government Printing Office, 2009. URL https://www.census.gov/content/dam/Census/library/publications/2010/acs/acs_design_methodology.pdf.

L. Vilhuber and K. McKinney. LEHD infrastructure files in the Census RDC - overview. Working Papers 14-26, Center for Economic Studies, U.S. Census Bureau, June 2014. URL https://ideas.repec.org/p/cen/wpaper/14-26.html.

D. Wagner and M. Layne. The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) record linkage software.

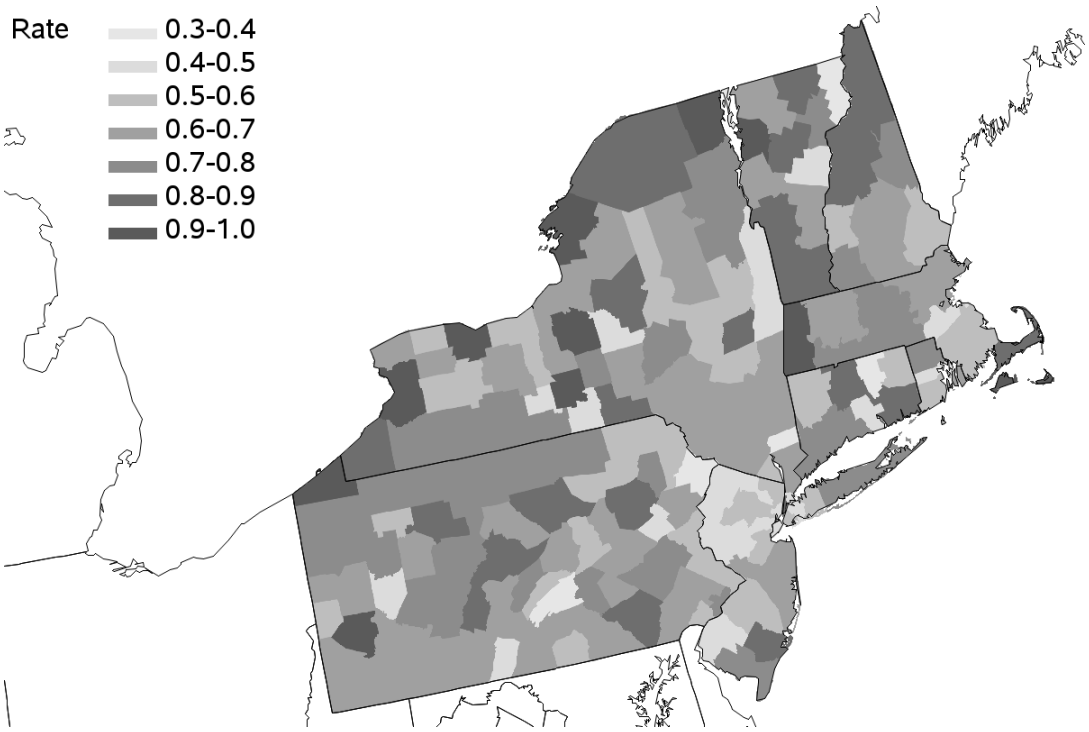
Working Papers 2014-01, Center for Administrative Records Research and Applications, U.S. Census Bureau, 2014. URL http://www.census.gov/srd/carra/CARRA_PVS_Record_Linkage.pdf.

N. Wasi and A. Flaaen. Record linkage using STATA: Pre-processing, linking and reviewing utilities. *The Stata Journal*, pages 1–15, 2015.

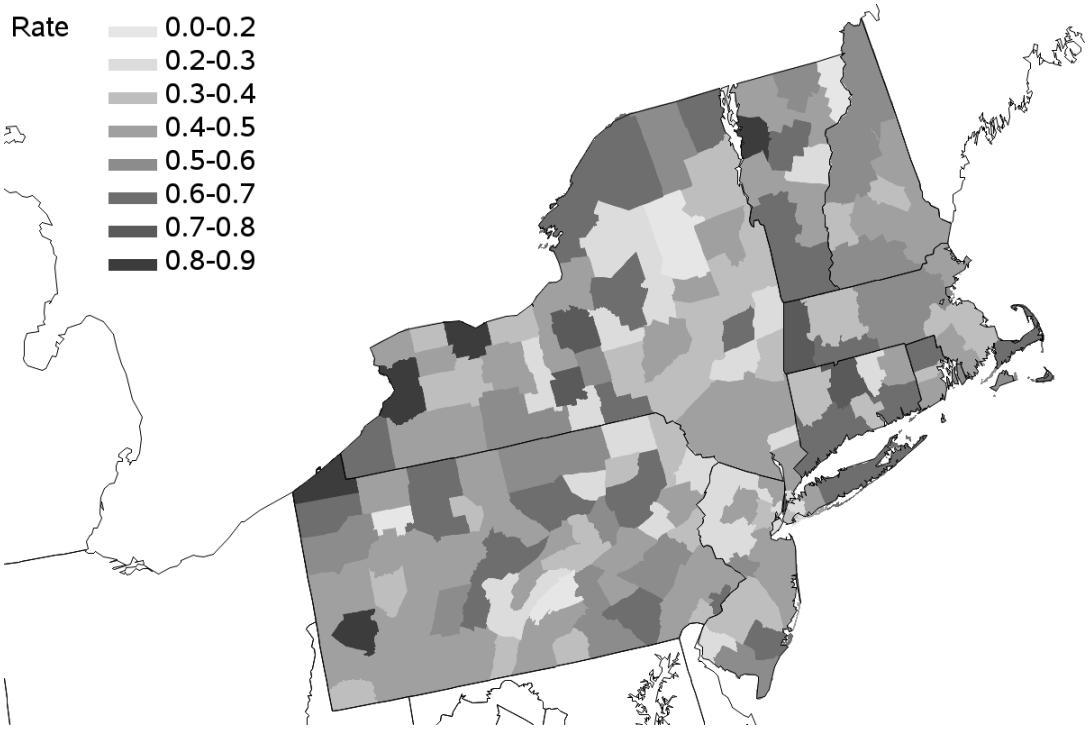
A Supplementary Tables & Figures

This section contains supplementary tables and figures. Where necessary a short description is included, otherwise we refer the reader to the text where the table or figure is referenced.

- **Figures A1**, **A2**, **A3**, and **A4** map the within-county commute shares for selected states already shown in Figure 2. Selections are for states in the Northeast, Midwest, South, and West, respectively.
- **Table A1** gives the correspondence between ACS and LEHD employment status using a narrower definition of LEHD employment than Table 4.
- **Table A2 & Table A3** elaborate on Table 3, breaking down the match rates to employers and establishments by characteristics of jobs, workers, and interviews and by industry, respectively.
- **Table A4** gives the average LEHD commute distance in miles by NAICS sector for the employer matched sample.
- **Table A5** provides the underlying numbers for Figures 4, 5, & 6.
- **Figure A5** gives the fraction of within-county commutes by distribution of commute distance. We provide it to give further detail to Figure 5 which simply shows the average within-county commute rate.
- **Figure A6** gives the average commute distance by commute distance bin. We provide it for further support to Figure 6.
- **Figure A7** shows the average commute distance by state firm size for multi-establishment firms. This is the complement to Figure 7, which is for single establishment firms.
- **Figure A8** shows the commute distance for the ACS and LEHD by number of possible establishments for the employer matched sample (as opposed to the establishment matched sample, in Figure 8).
- **Figure A9** gives the ACS and LEHD commute rate by number of possible establishments in the establishment matched sample. The LEHD commute is calculated using the Closest Establishment weights.
- **Figure A10** shows the commute distance for the ACS and LEHD by number of possible establishments. The figure uses the normed matcher weights to weight possible establishments. This is similar to the figures in Section 4.3. This figure keeps the same restrictions as Figure A8, but uses the matcher weights.



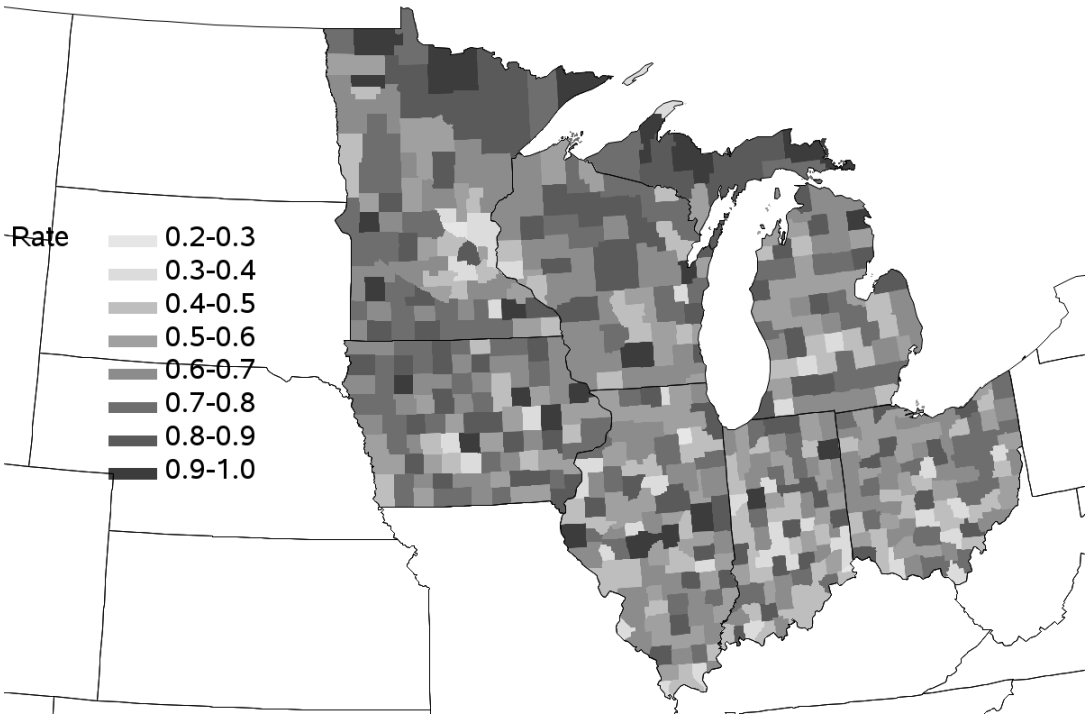
(a) ACS



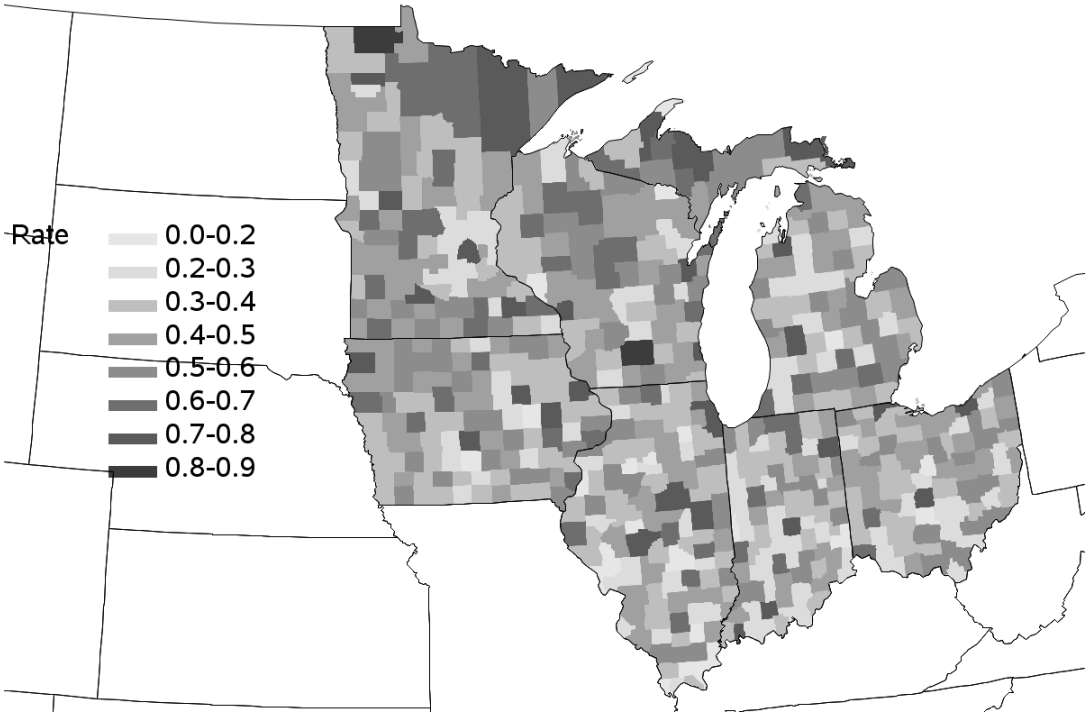
(b) LODES

Figure A1: Selected States in Northeast: Within-County of Residence Commute Rate

Notes: Shading coresponds to a larger share of county residents commuting to a workplace in the same county where they live, by decile bins from 0 to 1. See Table 1 for definitions of ACS and LODES public-use data. See Figure 2 for a map of the 48 contiguous states.



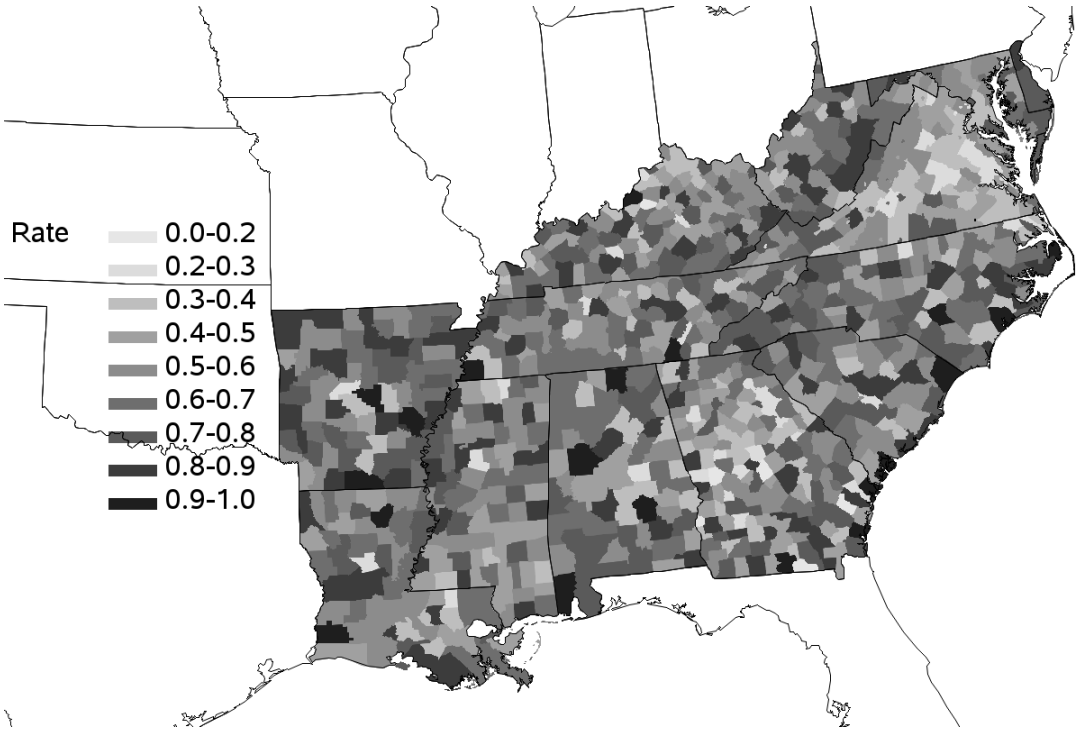
(a) ACS



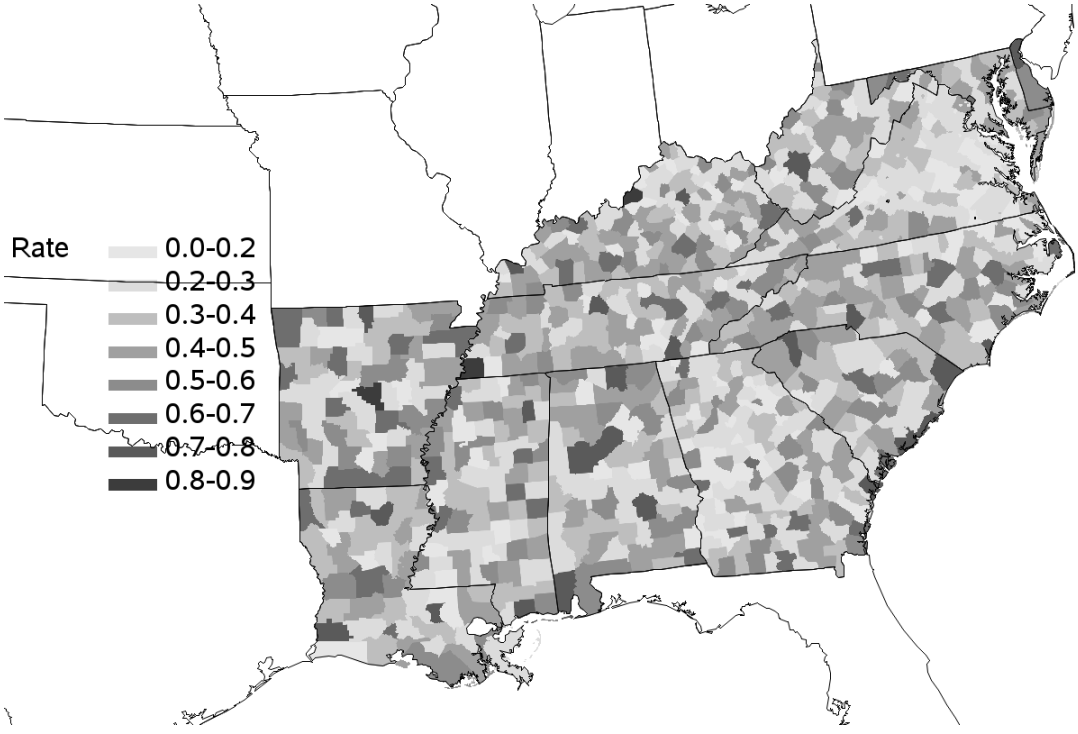
(b) LODES

Figure A2: Selected States in Midwest: Within-County of Residence Commute Rate

Notes: Shading coresponds to a larger share of county residents commuting to a workplace in the same county where they live, by decile bins from 0 to 1. See Table 1 for definitions of ACS and LODES public-use data. See Figure 2 for a map of the 48 contiguous states.



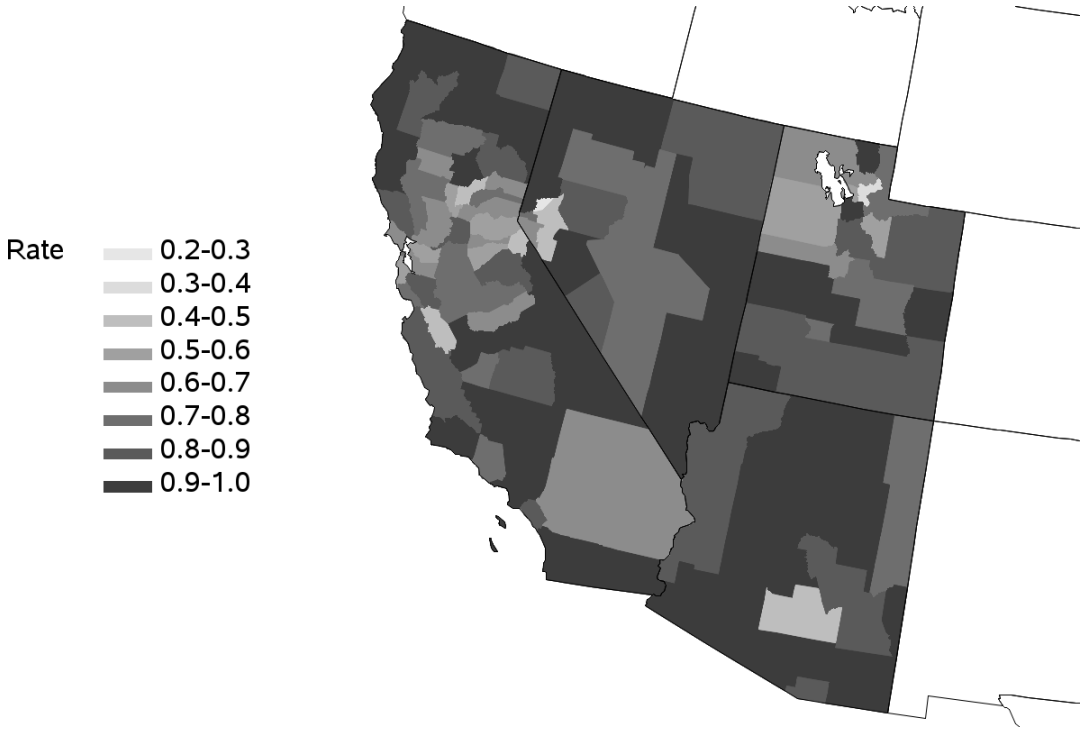
(a) ACS



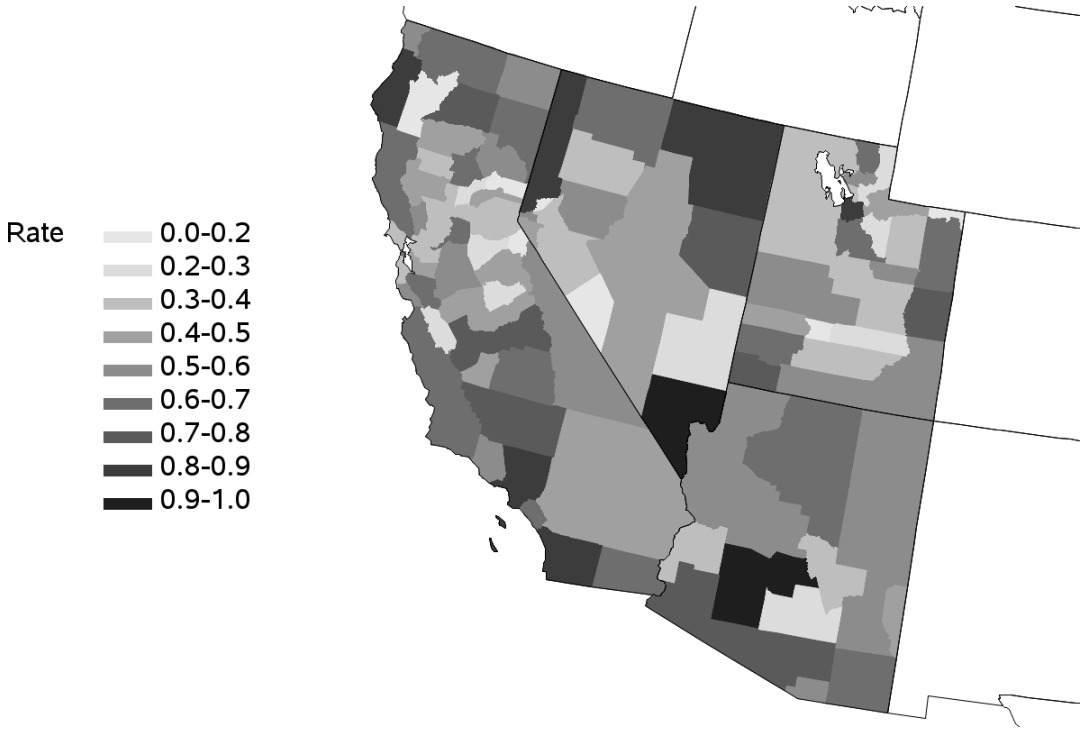
(b) LODES

Figure A3: Selected States in South: Within-County of Residence Commute Rate

Notes: Shading coresponds to a larger share of county residents commuting to a workplace in the same county where they live, by decile bins from 0 to 1. See Table 1 for definitions of ACS and LODES public-use data. See Figure 2 for a map of the 48 contiguous states.



(a) ACS



(b) LODES

Figure A4: Selected States in West: Within-County of Residence Commute Rate

Notes: Shading coresponds to a larger share of county residents commuting to a workplace in the same county where they live, by decile bins from 0 to 1. See Table 1 for definitions of ACS and LODES public-use data. See Figure 2 for a map of the 48 contiguous states.

Table A1: Joint Distribution of ACS and LEHD Employment Status, with 1-Quarter Overlap Window

	Link to LEHD in response quarter	
ACS	Employed	Not employed
Employed	41.8	4.9
Not employed	7.7	45.6

Notes: Sample defined in line 2 of Table 2, representing 667,000 ACS respondents. All numbers are cell percentages. LEHD jobs must have earnings in the ACS response quarter.

Table A2: Match Rates by Select ACS Characteristics

	(1) Employment Share	(2) Employer Match	(3) Establishment Match
Overall (% of Employed ACS)	100.0	72.60	36.74
Number of Establishments			
Single Establishment	38.18	100.0	49.15
Multiple Establishment	34.41	100.0	52.24
Employed in ACS, No Employer Match	27.40	00.00	00.00
Ownership			
Private, for profit	74.04	73.39	37.13
Private, not for profit	10.30	66.77	41.07
Local government	9.73	74.67	32.93
State government	5.92	69.41	30.67
Mode of Response			
Mail	72.56	74.92	41.56
CATI	10.39	31.12	24.65
CAPI	16.31	33.93	23.87
Person Number			
Person 1	53.35	74.38	40.19
Other	46.65	70.56	32.80
Urban/Rural			
Central city of MSA	27.85	70.12	36.01
Remainder of MSA	51.56	73.10	37.23
Outside MSA	20.50	74.68	36.51
Work From Home			
Car/truck/van	89.87	73.84	37.70
Walk	2.40	58.18	32.32
Worked at home	2.28	56.51	10.56
Other	5.45	65.13	33.92

Notes: $N = 311,000$. Sample contains all employed ACS respondents.

Table A3: Match Rates by NAICS Sector of those who have an employer match

	(1) Employment Share	(2) Employer Match	(3) Establishment Match
Overall (% of Employed ACS)	100.0	72.60	36.74
Agriculture, Forestry, Fishing, Hunting	0.96	49.30	18.62
Mining, Oil & Gas Extraction	0.57	74.51	22.68
Utilities	1.13	77.01	36.46
Construction	4.90	67.65	23.96
Manufacturing	12.39	77.72	42.28
Wholesale Trade	3.06	77.17	40.67
Retail Trade	11.72	79.54	39.71
Transportation and Warehousing	3.44	64.06	26.11
Information	2.40	70.41	38.98
Finance and Insurance	5.45	78.42	48.91
Real Estate and Rental and leasing	1.52	62.87	32.60
Professional, Scientific, & Technical Services	5.94	75.27	42.30
Management	0.08	82.69	54.62
Administrative Support & Waste Management	3.19	66.19	23.57
Education	11.66	74.29	36.10
Healthcare and Social Assistance	14.84	74.24	41.38
Arts, Entertainment, and Rec.	1.96	69.39	33.56
Accomodation and Food Services	6.17	70.19	32.32
Other Services, Except Public Admin.	4.09	48.88	26.60
Public Admin.	4.51	67.50	29.12

Notes: $N = 311,000$. Sample contains all employed ACS respondents. Column (1) shows employment share. Columns (2) and (3) show row percents.

Table A4: Average LEHD Commuting Distance by NAICS Sector

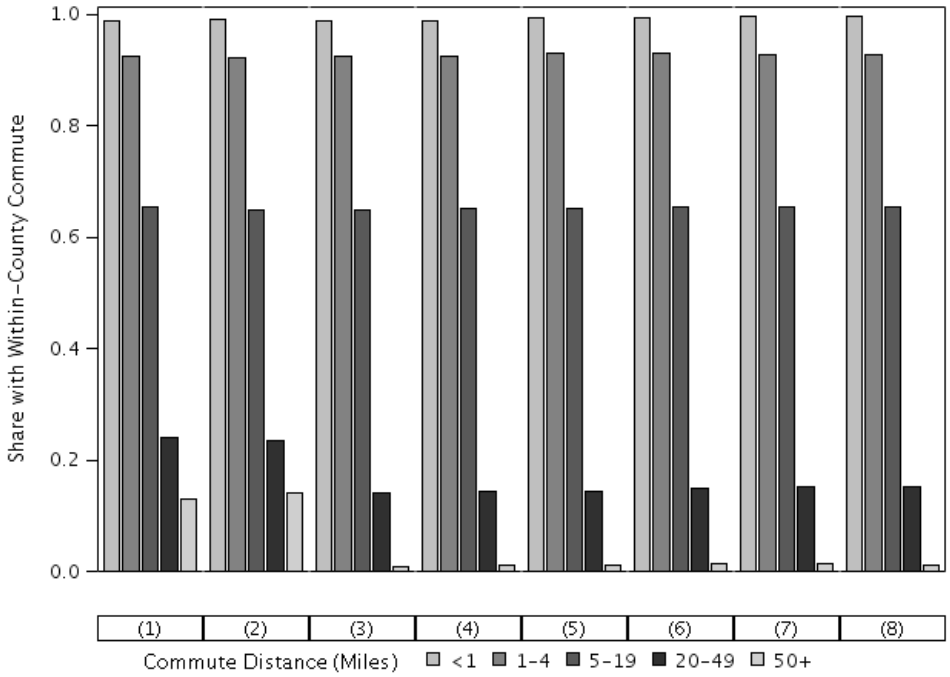
	Average Commute Distance (miles)
Agriculture, Forestry, Fishing, Hunting (11)	33.5
Mining, Oil/Gas Extraction (21)	62.5
Utilities (22)	30.5
Construction (23)	34.7
Manufacturing (31-33)	28.2
Wholesale Trade (42)	42.8
Retail Trade (44-45)	47.0
Transportation and Warehousing (48-49)	143.5
Information (51)	33.6
Finance and Insurance (52)	32.7
Real Estate and Rental and leasing (53)	39.3
Professional, Scientific, & Technical Services (54)	42.7
Management (55)	38.7
Administrative Support & Waste Management (56)	51.9
Education (61)	46.3
Healthcare and Social Assistance (62)	24.2
Arts, Entertainment, and Rec. (71)	29.0
Accomodation and Food Services (72)	34.3
Other Services, Except Public Admin. (81)	29.2
Public Admin. (92)	46.3

Notes: Employer Match sample $N = 226,000$. Average commute distance computed using Unit-to-Worker weights. Industry from LEHD job with highest probability of match.

Table A5: Underlying Numbers for Figures 4, 5, & 6

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Factors								
Weighting	LODES	LODES	LODES	LODES	LODES	**ACS**	ACS	ACS
Sample	Employer	**Estab.**	Estab.	Estab.	Estab.	Estab.	Estab.	**Employer**
Workplace	U2W	U2W	**Match**	Match	**ACS**	ACS	**ACS (edit)**	ACS (edit)
Residence	CPR	CPR	CPR	**ACS**	ACS	ACS	ACS	ACS
Figure 4: Average Commute Distance (miles)								
Average Commute Distance	25.0	18.8	10.9	9.8	9.8	9.8	9.8	10.1
10th percentile	1.8	1.5	1.2	1.2	1.2	1.1	0.6	0.6
90th percentile	46.7	35.5	22.2	21.5	21.5	21.6	21.7	22.3
Figure 5: Average Within-county Commute Rate								
	0.614	0.646	0.700	0.708	0.710	0.714	0.715	0.707
Figure 6: Distribution of Average Commute Distance (share)								
Miles								
< 1	0.049	0.061	0.081	0.082	0.083	0.085	0.124	0.121
$1 - 4$	0.237	0.262	0.321	0.324	0.324	0.325	0.289	0.282
$5 - 19$	0.447	0.460	0.475	0.477	0.477	0.474	0.468	0.472
$20 - 49$	0.175	0.157	0.107	0.106	0.106	0.105	0.108	0.113
$+50$	0.092	0.060	0.017	0.011	0.011	0.011	0.011	0.013

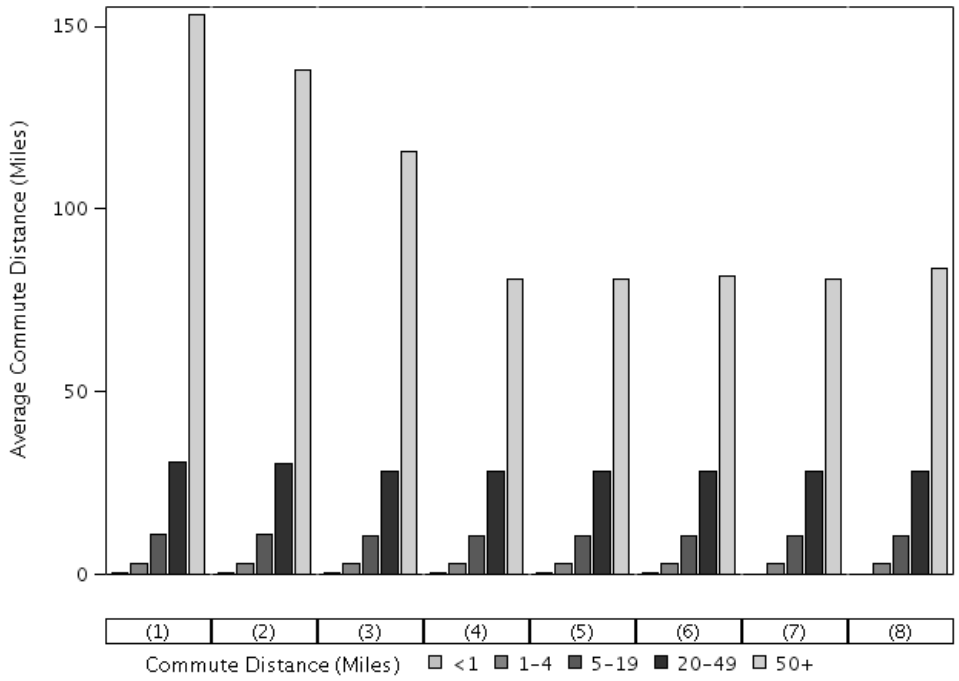
Notes: For description of numbered headers see Table 7 and text. For descriptions and graphic depictions of all numbers, see coresponding figures and the text. Due to rounding, shares for Figure 6 may not sum to one. Sample sizes are distance restricted for each stage and are as follows: column (1) 158,000, columns (2)-(6) 92,000, column (7) 83,000, and column (8) 135,000.



Factors	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weighting	LODES	LODES	LODES	LODES	LODES	**ACS**	ACS	ACS
Sample	Employer	**Estab.**	Estab.	Estab.	Estab.	Estab.	Estab.	**Employer**
Workplace	U2W	U2W	**Match**	Match	**ACS**	ACS	**ACS (edit)**	ACS (edit)
Residence	CPR	CPR	CPR	**ACS**	ACS	ACS	ACS	ACS

Figure A5: Fraction with Within-County Commute by Distribution of Commutes and Changes in Sample

Notes: Bars denote share of sample commuting to a workplace within the county of residence. See Table (7) and the text for a description of the different samples. For the underlying values, see Appendix Table A5.



Factors	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Weighting	LODES	LODES	LODES	LODES	LODES	**ACS**	ACS	ACS
Sample	Employer	**Estab.**	Estab.	Estab.	Estab.	Estab.	Estab.	**Employer**
Workplace	U2W	U2W	**Match**	Match	**ACS**	ACS	**ACS (edit)**	ACS (edit)
Residence	CPR	CPR	CPR	**ACS**	ACS	ACS	ACS	ACS

Figure A6: Average Commute Distance by Distribution of Commutes and Changes in Sample

Notes: Bars denote average commute distances for each sample in miles within each distance bin. Solid lines denote the 10th and 90th percentiles. See Table (7) and the text or a descrition of the different samples. For the underlying values, see Appendix Table A5.

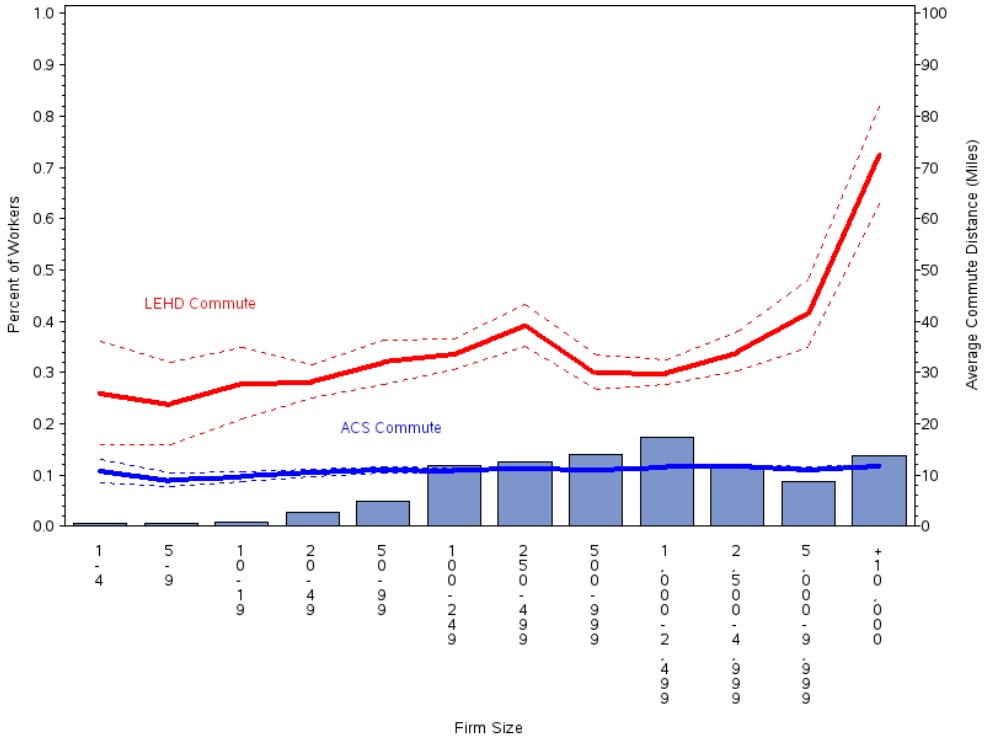


Figure A7: Commute Distance for Multi-Unit Employers by State Firm Size

Notes: $N = 64,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across state firm sizes corresponding to the left vertical axis. Solid lines denote average commute distance by state firm size corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is a subset of the *distance restricted Employer Match* sample, defined in 3. For this subset, ACS workers have multiple candidate establishments (the complement of those with a single establishment candidate), either due to matching to multiple employers or due to a matched employer having multiple establishments. ACS commutes use geocoded workplaces and LEHD commutes use U2W establishment probabilities.

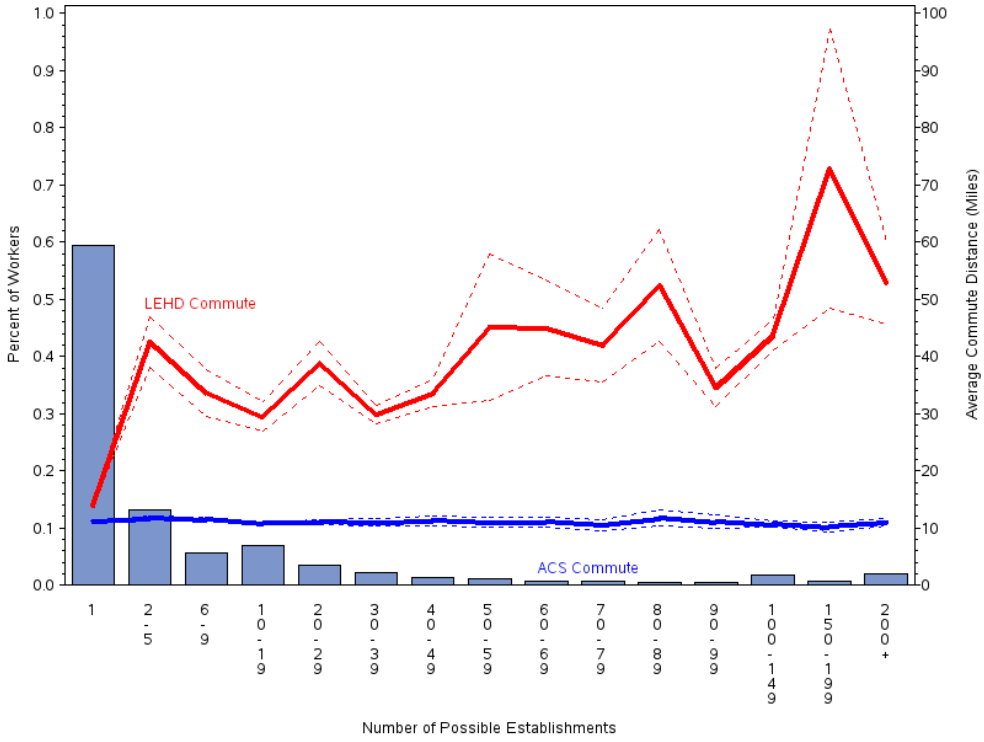


Figure A8: Commute Distance by Number of Establishment Candidates for Persons who Match to at Least One Employer with LEHD Commutes Using U2W Establishment Probability

Notes: $N = 158,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across number of candidate establishments corresponding to the left vertical axis. Solid lines denote average commute distance by possible establishment matches corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is the *distance restricted Employer Match* sample, defined in 3. ACS commutes use geocoded workplaces and LEHD commutes use U2W establishment probabilities.

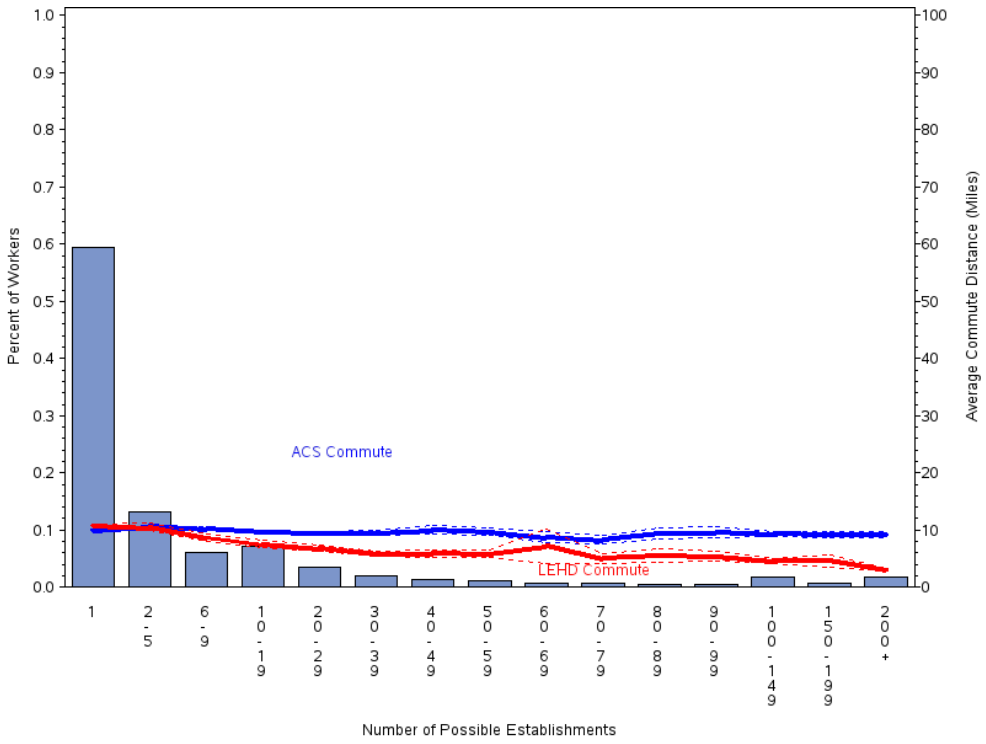


Figure A9: Average Commute Distance by Number of Establishment Candidates for Persons who Match to at Least One Establishment with LEHD Commutes Using Closest Establishment Probability

Notes: $N = 92,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across number of candidate establishments corresponding to the left vertical axis. Solid lines denote average commute distance by possible establishment matches corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is the *distance restricted Establishment Match* sample, defined in 3. ACS commutes use geocoded workplaces and LEHD commutes use establishments weighted with an indicator which evaluates to unity if the candidate establishment is closest to the residence.

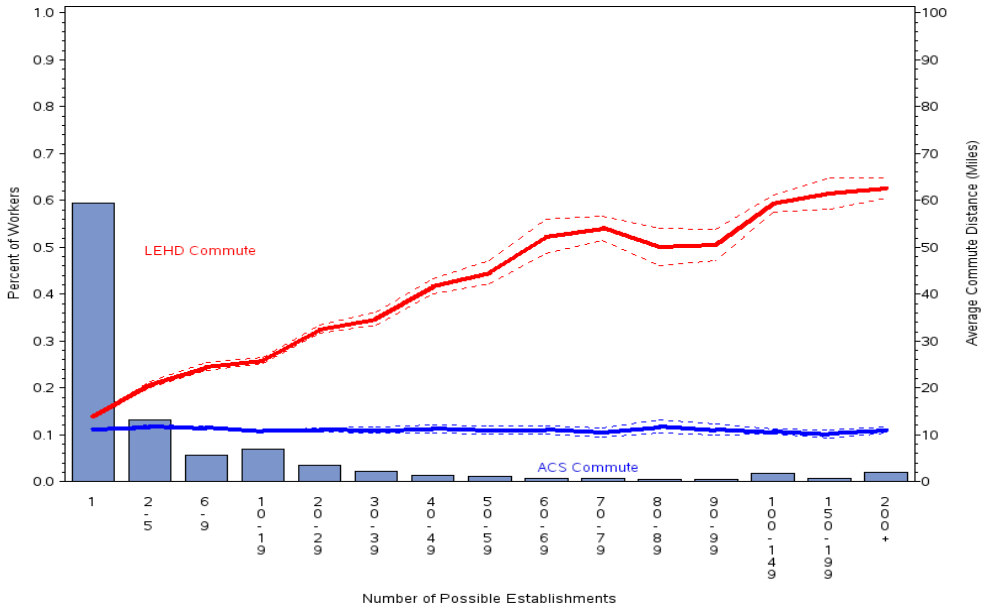


Figure A10: Average Commute Distance by Number of Establishment Candidates for Persons who Match to at Least One Employer with LEHD Commutes Using Normed Matcher Probability

Notes: $N = 158,000$. "LEHD Commute" is equivalent to d^L and "ACS Commute" is equivalent to d^A. Bars denote the distribution of workers across number of candidate establishments corresponding to the left vertical axis. Solid lines denote average commute distance by possible establishment matches corresponding to the right vertical axis. Dashed lines denote the 90% confidence interval of the mean. Analysis sample is the *distance restricted Employer Match* sample, defined in 3. ACS commutes use geocoded workplaces and LEHD commutes use establishment probabilities from the normed matcher model. The normed matcher weights use the probabilities from the matching model, but give some weight to all establishments, not just those deemed a match.