Developing a Residence Candidate File for Use With Employer-Employee Matched Data

Matthew R. Graham, Mark J. Kutzbach, and Danielle H. Sandler*
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233

May 5, 2016

1 Introduction

This paper describes the Longitudinal Employer-Household Dynamics (LEHD) program's efforts to use administrative records in a predictive model that describes residence locations for workers. This project was motivated by the discontinuation of a residence file produced elsewhere at the Census Bureau. The discontinued file provided only a single residence per person/year, even when contributing administrative data may have contained multiple residences. The goal of the Residence Candidate File (RCF) process is to provide the LEHD Infrastructure Files with residence information that maintains currency with the changing state of administrative sources and represents uncertainty in location as a probability distribution. This paper describes the motivation for the project, our proposed methodology, the administrative data sources, and the model estimation results. We find that the best prediction of the person-place model provides superior accuracy compared with previous methods and and performs well for workers in the LEHD jobs frame. Although our model predictions provide an indication of uncertainty, in expectation, the probability weighted model is less accurate than either the best prediction or the previous methodology. The paper outlines further work that may improve the representation of uncertainty and enhance the model with job timing and location information from the LEHD Infrastucture Files.

1.1 Background and Motivation

The LEHD program in the Center for Economic Studies (CES) at the Census Bureau uses job and employer information from states along with federal survey and administrative data to produce statistics on labor force dynamics including the Quarterly Workforce Indicators, LODES, and Job-to-Job Flows (Abowd et al. [2009]). States provide LEHD with quarterly files supplying the earnings of all workers covered by state unemployment insurance programs. These include state and local government employment as well as approximately 96% of all private sector wage and salary employment (Stevens [2007]). States also provide quarterly employer files listing establishment locations as well as industry, ownership, and size. LEHD combines these files into a Person History File, listing the earnings history of each job, and an Employer Characteristics File (ECF). LEHD also produces an Individual Characteristics File (ICF) based on federal survey and administrative data that provides demographic information on workers. LEHD uses these files to produce the public use datasets as well as for economic research.

The LEHD program requires place of residence information for several core processes, each of which expects a single, best residence for each worker. The Unit-to-Worker (U2W) imputation of establishments to persons uses residence to calculate implied commute distance from a workplace. The ICF imputes demographic characteristics based on the observed characteristics of neighbors. Additionally, the LEHD Origin-Destination Employment Statistics (LODES), disseminated through the OnTheMap web tool, tabulates the residence location of jobs in origin-destination tables and as a residence margin.

From its inception, the LEHD program has used the Composite Person Record (CPR) as a source of residence data. The CPR containes fields that provided a linkage between a unique person record and the location of a residential housing unit. The Center for Administrative Records Research and Administration (CARRA) used the Statistical Administrative Records System (StARS) to produce the CPR until the file was discontinued in 2011 (data year 2010).[1] CARRA delivered the MAF-ARF (Master Address File-Auxiliary Reference File) in place of the CPR in 2012 (data year 2011). The MAF-ARF was found to differ from the CPR in a number of ways, including a difference in coverage and a lack of deduplication among PIKs. LEHD was able to produce a deduplicated version of the MAF-ARF by defining some very basic business rules that were implemented by the CES Data Staff.

*DISCLAIMER: 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.

[1] See https://www.census.gov/content/dam/Census/about/about-the-bureau/SORNs/CEN-08.pdf

The 2012 process of unduplicating the MAF-ARF was not sustainable, and in creating a process internal to LEHD, we were able to address quality and suitability issues specific to the needs of the program. As a first step to developing a permanent replacement for the CPR/MAF-ARF, the LEHD program developed a process (RCFv0.5) that replicated the methodology of the MAF-ARF. LEHD used the output of this new process to supply the 2013-2015 cycles of LODES processing (data years 2012-2014) with residential information. The iteration described here (RCFv1.0) develops a predictive model for residence location that is customized to the needs of employer-employee matched data. The RCF file contains PIKs and a weighted, preferred list of residential locations geocoded to 2010 census tabulation blocks along with metadata on geocoding outcomes. Integration with the LEHD Infrastructure Files is scheduled as a next step.

2 Methodology

The production specification for the RCF is to transform a set of annual administrative source files listing person/location into a file with preference weights for each person/location and with no remaining source information. Construction of the RCF does not alter the fundamental principle of how LEHD uses residence data. Namely, LEHD processes requiring a place of residence still access a composite file, rather than the source files themselves. Downstream processes currently expect a single residence for each worker in a year, but imputations and tabulations could be modified to consider a set of residences with model-determined preference weights. Neither the CPR nor the RCF indicate which source provided each record. The present analysis considers a baseline, rank-order model that is analagous to the CPR methodology as well as several specifications of a person-place model that offers more customization and provide more information on the uncertainty of locations for a person.

2.1 Rank-Order Model

The CPR limits residence data to a "best" record for each person in a year.[2] Administrative data handlers chose a best record by developing a source priority order. The LEHD program has little information on the development of the priority order for residence data in earlier years. For processing the 2010 MAF-ARF residence data, CES used a priority order based on findings from the 2010 Census Match Study [Rastogi et al., 2012]. The study linked responses to the 2010 Decennial Census with administrative records to identify the sources that corresponded best with a person's response location on the census reference date, April 1, 2010. (The April 1 date is useful because it is coincident with the LEHD Beginning-of-Quarter employment definition for Quarter 2, used for the LODES snapshot of jobs.) CES deduplicated the MAF-ARF according to the priority order and retained the highest priority residence source for each person.

There are several drawbacks with using a priority order to create a deduplicated series of addresses for LEHD processes:

- First, deduplication disposes of information on the distribution of possible residential locations that may apply to a worker at a point in time. For some other missing data problems, LEHD reflects uncertainty by the tabulating statistics with weights associated with the probability of any piece of information. Furthermore, a worker may have multiple residences over a period of time, each of which might be associated with a different job. Prioritizing one address over another degrades the interpretation of residence as the home of a worker while employed at a particular job.
- Second, the priority order developed in the 2010 Census Matching Study is based on an April 1 reference date for a single year, but LEHD produces a quarterly series of job statistics beginning in 1985 for some states. Seasonal or longitudinal variation in the quality of the sources might alter the priority order.
- Third, the availability of sources varies longitudinally and not all sources included in the study are available for the RCF. The priority order of the sources might change if it were limited to those available for the RCF in each year.
- Fourth, this priority list does not reflect the relative strength of each source for different populations. Although the 2010 match study used demographic information in its predictive model, we do not have disaggregated priority lists.

We compare this baseline methodology (RCFv0.5) with proposals for a person-place methodology (RCFv1.0) described below.

2.2 Person-Place Model

As a basis for the RCF, we build upon a two stage process that was originally developed by David Brown at the Census Bureau in order to create a model for estimating the occupancy of households that did not respond to the 2010 Decennial Census [Brown, 2013]. Brown [2013] uses responses to the 2010 Census as a truth set for training and validating the model.[3] Brown [2013] uses a two-stage logistic model of location agreement. The first stage estimates an equation for each adminstrative data source and the second stage, pooling residences from all sources, makes use both of indicators for the presence of each source as well as predicted probabilities from the first stage. These probabilities give the expected validity of each source for each person. The predicted

[2] As an exception, the 2001 CPR included multiple residences for some persons.

[3] Other studies at the Census Bureau that trains and validates an administrative data model using the 2010 Census include [Rastogi et al., 2012] and [Steeg Morris et al., 2016, forthcoming]

probabilities from this second stage are used to create preference weights for each person at each reported residence. We build on and adapt this model in several respects, described in more technical detail below.

First, given that LEHD produces quarterly jobs data, it is necessary to train the residence model with greater frequency than is possible with the Decennial Census. We use the American Community Survey (ACS), a continuous household survey with national coverage since 2003. Both the American Community Survey (ACS) and Decennial Census require respondents to reside at an address in the mailing frame, which provides a concept of residence that is fairly consistent across time, geography, and populations. Using the ACS allows us to train and validate the model independently for any given year. For this study, we implement the model for a single study year.

Second, to make the residence file more robust to the gaps in the reporting of administrative sources for a person in any year, we supplement the residence list with addresses appearing in prior and subsequent years in the model. Lagged or later residences may have less predictive power for where someone is in the study year, so along with including these records, we add parameters that capture the longitudinal history of reporting for each source. This model-based solution will be more informed than the longiduinal edits currenly used in LEHD processing.

Lastly, we limit the model estimation to working-age persons ever appearing in LEHD job histories and use demographic information already included the LEHD Infrastructure Files. We also evaluate the model specifically for persons employed in LEHD during the study year.

2.3 Model Description

We first make eligibility restrictions on both the survey and administrative residence files based on address reference dates, person characteristics, and the availability of linking information. We use identifiers encoding personal identifying information to link residence candidates from the administrative sources to a person's survey based residence location. We designate one portion of this linked set as a training sample and the remainder as a validation sample. Note that in this case, the vast majority of administrative records eligible for inclusion in the residence file do not link to any person in the survey file.

For individual i with location l from source s, we specify an agreement, Γ_{ils} as

$$\Gamma_{ils} = I(Survey_{il} = Admin_{ils}) \tag{1}$$

where I is an indicator function for agreement of the residence location between the person's survey response and administrative source at some level of geography. The model could be specified for a range of geographic precision for Γ_{ils}, including by address identifier, Census tabulation geography, county, or state. For the present analysis, we use Census block, the most detailed geographic tabulation of residence published by LEHD in LODES.

Starting with the training sample, we explain the variation of this binary agreement variable with a logistic model estimated separately for each of the S sources. We specify the model as

$$\Gamma_{ils} = \frac{exp(\alpha_s + \beta_s X_{ils})}{1 + exp(\alpha_s + \beta_s X_{ils})} \tag{2}$$

where X_{ils} is a vector of individual and source characteristics including demographic information as well as the reference date of the source. For the same observations, we predict $\hat{\Gamma}_{ils}$ given our estimates of $\hat{\alpha}_s$ and $\hat{\beta}_s$ as well as the characteristics X_{ils}. We use these expected values for each person/location/source in the second stage.

For the second stage, we deduplicate the data by person/location (collapsing cases of multiple sources for the same person/location). We add indicator variables S_{il} and the predicted probabilities $\hat{\Gamma}_{ils}$ for each source appearing for that person/location. We then estimate a second logistic model, specified as

$$\Gamma_{il} = \frac{exp(\gamma + \sum_{s=1}^{S}(\phi_{ils} S_{il} + \lambda_{ils} S_{il} \hat{\Gamma}_{ils}))}{1 + exp(\gamma + \sum_{s=1}^{S}(\phi_{ils} S_{il} + \lambda_{ils} S_{il} \hat{\Gamma}_{ils}))} \tag{3}$$

Note that the predicted probabilities are set to zero in the case where there is no corresponding source, so we only write them as an interaction with the source indicators.

Turning to the validation sample, we apply the parameter estimates from equations (2) and (3) to compute expected agreement for each person/location, where

$$\hat{\Gamma}_{il} = \frac{exp(\hat{\gamma} + \sum_{s=1}^{S}(\hat{\phi}_{ils} S_{il} + \hat{\lambda}_{ils} S_{il} \hat{\Gamma}_{ils}))}{1 + exp(\hat{\gamma} + \sum_{s=1}^{S}(\hat{\phi}_{ils} S_{il} + \hat{\lambda}_{ils} S_{il} \hat{\Gamma}_{ils}))}$$

We then compute preference weights for each individual i. For a person with L_i locations across all sources, we create an aggregated value of the predicted probabilities for the denominator of the weight. For each location, we use the predicted probability for the person/location for the numerator, such that each person/location is assigned a weight, W_{il}, specified as

$$W_{il'} = \frac{\hat{\Gamma}_{il'}}{\sum_{l=1}^{L_i}(\hat{\Gamma}_{il})} \tag{4}$$

The same methodology used to calculate preference weights for the validation sample is then applied to the entire frame of persons with administraive residence data. The resulting RCF file, which is annual at this stage, gives preference weights for each person/location and retains no source indicators.

3 Data

This study focuses on the year 2012 for presentation of the data infrastructure and model estimation.

3.1 Administrative data sources

This project uses administrative data on residence location from the Internal Revenue Service (IRS), the Department of Housing and Urban Development (HUD), the Department of Health and Human Services (HHS), the Selective Service System (SSS), and the U.S. Postal Service (USPS). The RCF uses the following administrative source files:

IRS 1040 Individual Tax Returns (1040/IMF)

IRS 1099 Information Returns Master File (1099-R/IRMF)

HUD PIC Multi-Family Tenant Characteristics System / PIH Information Center (MTCS/PIC)

HUD TRACS Tenant Rental Assistance Certification System (TRACS)

HUD CHUMS Computerized Home Underwriting Management System (CHUMS)

HHS IHS Indian Health Service - Patient Registration

HHS CMS Medicare Enrollment Database - 100 percent Production File

SSS Selective Service System Registration Files

USPS NCOA National Change of Address File (NCOA)

Detailed descriptions of each data source are available in the appendix.

3.2 Definitions

The RCF is meant to be integrated with jobs data defined by earnings of a person at a job in a quarter. The LEHD earnings files identify job holders with a Protected Identification Key (PIK). Likewise, the administrative source files identify persons with a PIK. The Census Bureau maps administrative and survey data records to a PIK using personal identifying information Wagner and Layne [2014]. In cases where records could not be mapped to a PIK, data integration is not possible.

We use the Master Address File ID (MAFID) to define a residence location, an identifier used throughout demographic survey areas at the Census Bureau. The MAF is the residence frame for both the Decennial Census and the American Community Survey. CARRA has already geocoded address fields in administrative source files to MAFIDs, where possible.

As with the CPR, the RCF defines the period of residence based on the reference data of a source file. Each source file has an independent schedule of when it is collected, produced, and delivered to the Census Bureau. Where possible, the RCF uses reference date fields to define the year of residence. In the absence of a reference date, the RCF uses metadata on the origin of the file to infer, at a minimum, the year of the residence records.

3.3 Source Summary

Table 1: Administrative records with a MAFID, 2012

Address source (detailed)	Records (millions)
IRS 1040	262
IRS 1099	624
HUD PIC	7
HUD TRACS	2
IHS	5
Medicare	48
SSS	15
NCOA	37
Total	1,000

When reading in the source files, we only retained records with a PIK. Of those, 85 percent overall had a MAFID, almost all of which had complete 2010 Tabulation Geography. Some sources were notable in having a lower percentage of records with valid MAFIDs. In 2012, these include IHS (73 percent) and IRS 1099 weeks 42-52 (77 percent). Table 1 lists the totals of PIKed records with a MAFID before any deduplication within or between sources.

Table 2: Source availability for PIKs, 2012

Address source (detailed)	Percent with source
IRS 1040	85.3
IRS 1099	70.3
HUD PIC	2.2
HUD TRACS	0.8
IHS	0.9
Medicare	16.4
SSS	5.0
NCOA	4.8

Table 2 gives the share of all PIKs that have at least one residence record provided by a source. Almost all PIKs have an IRS sourced address available, with 85 percent having a 1040 address and 70 percent having a 1099 address. Availability of the other sources is fairly consistent with the scope of the programs providing data.

We produce residence frames for 2011, 2012, and 2013 consisting of the set of unique source/address records per PIK, where each residence has a valid MAFID. Table 3 lists the record count for the 2012 file. Over 99% of these records have residence information precise enough to be geocoded to a Census block.[4]

Table 3: Record counts in millions, 2012

Record Type	Count (Millions)
Unique person/source/MAFID	593
Unique person/MAFID	376
Unique person	296

After applying the priority rules similar to the CPR to deduplicate by PIK, we retain 296 million records for 2012 (and a similar count for the other years). These totals are in line with the 2010 Census Match Study, which found 302 million records with a PIK and MAFID using both federal and commercial source data.

3.4 Training and Validation Sample

To train the predictive residence model, we draw a sample of 5.3 million respondents from the ACS in 2012. Because we are focused on residence prediction for a jobs frame, we first limit the sample to the 5.0 million respondents who can be linked to a PIK. The PIK rate for the ACS is 94.0%. The only information we retain from the ACS is the quarter of response within 2012 and the place of residence Census block.

We further limit the sample to those persons recorded as ever having worked in LEHD, or the set of persons in the ICF. Although the ACS includes demographic information, we obtain variables for age, sex, race, ethnicity, educational attainment, and native birth from the ICF, which is based on NUMIDENT, 2000 Census responses, and imputations. We calculate age from the ICF at the response date to ACS and only retain persons from aged 14 to 74. Limiting the sample to this set of likely workers improves the applicability of the model estimates to the frame of LEHD workers. We partition this set of 3.4 million persons and their associated records randomly to the 70 percent training sample and 30 percent validation sample.

As is described in the methodology, we then link the ACS file with the administrative residence data described above. We use data from 2011 to 2013 as potential address matches for 2012 ACS responses. For just 2012, the merging resulting in 21.5 million unique PIK/source/MAFID records and 10.0 million unique PIK/MAFID records. The first stage of the model will draw from the first quantity while the second stage will draw from the latter.

Table 4 shows the number of MAFIDs available per PIK, both in 2012 and for all three years. While 72.2% of PIKs have one and only one MAFID from our source files in 2012, 21.1% have two or more potential addresses. These are the individuals for which we need a deduplication methodology. We only use persons with a link to at least one administrative residence record for the model estimation, but retain the remaining 6.2% for overall evaluation of the model in linking residence locations to persons or workers. When we expand the set of candidate records to include locations from sources in a lagged or later year, the share of persons with multiple MAFIDs rises to 38.1% and the share with no MAFID falls to 2.6%.

[4]For the present analysis, we retain this small set of records with a missing or incomplete geocode and code them as lacking agreement with the survey location.

Table 4: Availability of administrative address for PIKs (quantities in thousands)

MAFIDs per PIK	Count 2012	Percent 2012	Count all years	Percent all years
0	213	6.18	91	2.64
1	2,503	72.72	2,039	59.24
2	575	16.71	766	22.25
3	122	3.53	320	9.30
4	24	0.70	135	3.93
5+	5	0.15	91	2.64
Total	3,442	100.00	3,442	100.00

Table 5 shows the longitudinal history trend of addresses by source file. The header gives a triad that indicates whether the address was present in 2011, 2012, and 2013. So, if an address is present in 2011 in the Medicare file for a individual, but is not present for 2012 or 2013 in the Medicare file, then that individual would have a triad value of "100", a pattern for 5.5 percent of addresses linked to the sample from the Medicare file. Overall, almost half of all addresses appearing in the three year window occur in all three years (111). The least common pattern is a one year hole (101), which could be due to either a temporary move, a gap in the administrative data for the person, or geographic measurement error in the administrative data. Some of the patterns reflect the nature of the source files. For example, the NCOA file, based on change of address, has very few records with a PIK/MAFID in all three years, but many one year records.

Table 5: Longitudinal residence histories (triads), pattern percentages by source

Source	PIK/MAFID/Sources (thousands)	000	100	010	001	110	101	011	111
None	91	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IRS1040	3,947	0.0	13.5	5.9	11.8	7.6	1.7	8.2	51.5
IRS1099	4,555	0.0	12.3	5.6	13.4	7.9	1.8	10.2	48.8
HUDPIC	78	0.0	14.3	5.9	17.9	12.3	2.8	13.2	33.6
HUDTRACS	25	0.0	14.9	8.5	18.9	12.3	0.8	10.4	34.2
IHS	74	0.0	9.5	4.1	8.7	4.4	2.8	9.8	60.8
Medicare	641	0.0	5.5	2.3	16.6	4.0	0.9	12.6	58.1
SSS	227	0.0	11.6	1.5	12.1	9.8	0.5	14.0	50.5
NCOA	403	0.0	40.7	34.0	13.3	7.6	1.5	2.3	0.6
Total	10,041	0.9	13.3	6.5	12.8	7.5	1.6	9.3	48.0

4 Model Analysis

This section summarizes the model estimation and evaluation.

4.1 Specifications

We execute both the baseline, rank-order methodology and the person-place methodology described in Section 2. For each model, we use agreement by residence Census block as the outcome variable to be explained. In order to better understand the contribution of various parameters to the person-place model relative to the rank-order model, we estimate and evaluate several person-place model specifications across different sets of source years:

1. Rank-Order model (2012 MAFIDs)
2. Person-place model (2012 MAFIDs)
3. Person-place model (2011-2013 MAFIDs)
4. Person-place model with first stage triads (2011-2013 MAFIDs)
5. Person-place model with first and second stage triads (2011-2013 MAFIDs)

The first model is simply the application of the source rank-order rule to the administrative sources. All of the person-place models include the demographic variables from the ICF and the response quarter indicators from the ACS. The second item is a person-place model estimated on the same record set as the rank-order model. The third model is the same as the second, but with the addition of lagged and later year source files. The fourth model adds indicators for the longitudinal pattens (triads) of each source to the first stage as additional characteristics. The fifth model not only includes the triads in the first stage, but also adds

indicators for each source/triad in the second stage as well as predicted probabilities for each source/triad. Considering Table 5, the fourth model adds a parameter to the first stage for each triad column (excluding "000" and "111"). The fifth model adds those in the first stage, but also adds parameters for each cell in the table to the second stage. Thus, the fifth model makes the most use of longitudinal information.

4.2 Estimation

We present results for estimating the fourth model listed above (person-place with longitudinal triads in the first stage) on the 70 percent training sample. This model explains Census block agreement with our ACS extract from 2012 using MAFIDs from 2011 to 2013. Table 6 gives the log odds ratios for the first stage regression, as discussed in Section 2, for a single source, the IRS 1040. Reviewing the log odds ratios, all of the longitudinal triad estimates are less than one, indicating that sources not appearing in all three years (the omitted class) are inferior predictors of location. The reference quarter variables indicate worse correspondence for ACS respondents in the 4nd quarter, relative to the first quarter (as shown by the log odds ratio of less than one). The match is more accurate for middle aged responents than younger respondents (64 to 74 omitted), with those aged 18 to 24 having the worst match rate. White and Asian respondents, non-Hispanics, and those with higher educational attainment tend to have better address correspondence.[5]

Table 6: First stage person-place estimation for IRS 1040 addresses with first stage triads

Variable	Log odds ratio w.r.t. omitted	95% C.I. lower bound	95% C.I. upper bound
001 vs 111	0.05	0.05	0.05
010 vs 111	0.13	0.13	0.13
011 vs 111	0.38	0.38	0.39
100 vs 111	0.03	0.03	0.03
101 vs 111	0.29	0.28	0.29
110 vs 111	0.23	0.23	0.23
ACS Quarter 2	0.99	0.98	1.00
ACS Quarter 3	0.98	0.98	0.99
ACS Quarter 4	0.95	0.94	0.96
Age <18	0.64	0.63	0.65
Age 18 to 24	0.47	0.46	0.47
Age 25 to 34	0.65	0.64	0.66
Age 35 to 44	0.84	0.83	0.85
Age 45 to 54	0.97	0.96	0.99
Age 55 to 64	0.99	0.98	1.01
Female	0.97	0.96	0.98
Black	1.04	1.03	1.05
AIAN	0.46	0.45	0.47
Asian	1.10	1.08	1.12
NHPI	1.17	1.09	1.25
Two or more	1.02	1.00	1.05
Hispanic	1.02	1.01	1.03
High School	1.03	1.02	1.04
Some college	1.05	1.04	1.06
Bachelor's +	1.08	1.07	1.10
Native born	0.86	0.85	0.87

Table 7 gives the log odds ratios for the second stage regression (also of the fourth model listed above), which includes indicator variables for each of the source files and the probability weights from the first stage. While the log odds ratios provide some indication of the correspondence of the sources with ACS residences, interpretation would require considering both the indicator and interaction effect of each source. While some sources, such as IRS 1040 have strong effects for both the indicator and interaction variables, others, such as SSS and HUDPIC are most accurate for sub-populations identified in the interaction effect.

4.3 Evaluation

Having estimated the person-place model on the 70 percent training sample, we now apply the model to make predictions for the 30 percent validation sample. For evaluation purposes, we define a weighted agreement rate for N persons in the validation sample

[5] Similar tables exist for the other source files as well. Since there is no causal interpretation to these tables, we only choose to show this one for illustrative purposes.

Table 7: Second stage person-place estimation with first stage triads

Variable	Log odds ratio w.r.t. omitted	95% C.I. lower bound	95% C.I. upper bound
IRS1040	0.33	0.33	0.33
IRS1040*Prob	46.39	45.79	47.01
IRS1099	0.48	0.48	0.49
IRS1099*Prob	13.55	13.38	13.73
HUDPIC	0.21	0.20	0.23
HUDPIC*Prob	110.58	99.84	122.47
HUDTRACS	0.36	0.32	0.39
HUDTRACS*Prob	57.45	48.96	67.42
IHS	0.26	0.24	0.28
IHS*Prob	4.77	3.58	6.36
Medicare	0.50	0.48	0.52
Medicare*Prob	4.53	4.32	4.75
SSS	0.13	0.12	0.14
SSS*Prob	39.11	36.08	42.39
NCOA	0.16	0.15	0.17
NCOA*Prob	25.25	23.47	27.18

as

$$\text{Agreement Rate} = 100 * [\frac{1}{N}\sum_{i=1}^{N} I_{L_i>0}(\frac{1}{L_i}\sum_{l=1}^{L_i}\Gamma_{il}W_{il})]. \tag{5}$$

We include in N all persons in the validation sample, even if they have no candidate MAFID. The indicator function $I_{L_i>0}$ designates those persons with no MAFID as contributing zero to the match rate. We present the agreement rate in percentage terms.

For the full ACS sample (with no employment restriction), Table 8 gives the agreement rates for the baseline, rank-order model and each of the person-place specifications. For each person-place model, we present the agreement rate under two weighting assumptions. First, we set the weight W_{il} from equation (4) equal to one for the MAFID with the highest $\hat{\Gamma}_{il}$. We term this weight as the "Best" prediction. Second we report estimates based on weights from equation (4). We term this weight as the "Weighted" prediction. We compare these with several bounding models in the second panel. The upper bound of the agreement rate is dictated by the share of persons in the sample with at least one MAFID, which we compute both for 2012 and all years (2011-2013). Within those bounds, some persons may have no MAFID that agrees with the ACS residence. As an upper bound to the modeling predictions, we compute the maximum agreement rate assuming that $W_{il} = \Gamma_{il}|L_i > 0$ (so that the weight is one for an agreeing MAFID and zero otherwise). We also report an uninformed agreement rate, placing equal weight on each MAFID with $W_{il} = 1/L_i|L_i > 0$.

Table 8: Model comparison, full ACS validation sample

Preference weight	Rank-Order	PP-One Year	PP-Multi-Year	PP-Multi-Year	PP-Multi-Year
Best prediction	78.4	78.5	79.3	81.8	81.9
Weighted prediction	N.A.	76.0	74.2	76.6	76.8
Has MAFID (any year)			97.4	97.4	97.4
Has MAFID in 2012	93.8	93.8			
Perfect matching	82.4	82.4	89.2	89.2	89.2
Uninformed matching	74.6	74.6	71.9	71.9	71.9
Longitudinal History Triads				1st stage	1st & 2nd stage
ACS Observations	1,033,000	1,033,000	1,033,000	1,033,000	1,033,000

Using information only from the reference year, the most basic person-place model has a Best prediction agreement rate of 78.5 (in column two), only slightly higher than the rank-order rate of 78.4 (in column one). Both of these are higher than the Weighted rate of 76.0. Note that both models fall between uninformed and perfect rates of 74.6 and 82.4. The lack of substantial improvement for the person-place model relative to the rank-order model suggests that the relative importance of the various sources has not changed much from when the source order was determined and that there is relatively little gained from using demographic characteristics in the first stage of the person-place model. The relative contribution of this model might improve with the inclusion of additional characteristics, if the sources changed in nature, or of the set of sources were updated. The relatively worse performance of the weighted model suggests that there may be some mis-specification of the model, resulting in the predicted order being more reliable than the exact value of each prediction.

Expanding the set of candidate location/sources to 2011 and 2013 substantially improves the performance of the person-place model relative to the rank-order model. As can be seen from Table 4, expanding the record set introduces many more cases of multiple MAFIDs. This complexity could potentially make modeling more difficult, as is apparent from the drop in the agreement rate for uninformed matching to 71.9. However, it also provides more opportunities to identify the ACS response location, as is apparent from the rise in the agreement rate for perfect matching to 89.2. Simply adding these sources to the person-place model, in the third column, increases the agreement rate to 79.3.

Adding the first-stage longitudinal triads, in the fourth column, further increases the Best prediction to 81.8. Thus, expanding to additional years improves the agreement rate, but the inclusion of the longitudinal history triads is important for appropriately discounting the contribution of those years relative to the reference year. The final person-place model adds more complexity, with additional history parameters for each source in the second stage, but these have only a minor contribution. The agreement rate of the Best and Weighted predictions rise from 81.8 to 81.9 and 76.6 to 76.8.

Because the goal of this project is to predict administrative residences for workers in employee-employer matched data, we next restrict the ACS sample to those with at least one LEHD job in the year they responded. We present the results for employed respondents in Table 9.[6] For this worker sample, we find the same pattern as before, but with higher agreement rates. The highest rate we achieve is for the last person-place model, with an agreement rate of 83.4 (in column five) compared to uninformed and perfect rates of 72.0 and 91.3, as well as a rank-order rate of 80.3 (in column one).

Table 9: Model comparison, ACS validation sample employed in LEHD

Preference weight	Rank-Order	PP-One Year	PP-Multi-Year	PP-Multi-Year	PP-Multi-Year
Best prediction	80.3	80.4	80.7	83.2	83.4
Weighted prediction	N.A.	77.5	75.2	77.9	78.2
Has MAFID (any year)			98.5	98.5	98.5
Has MAFID in 2012	95.8	95.8			
Perfect matching	84.5	84.5	91.3	91.3	91.3
Uninformed matching	75.6	75.6	72.0	72.0	72.0
Longitudinal History Triads				1st stage	1st & 2nd stage
ACS Observations	659,000	659,000	659,000	659,000	659,000

4.4 Release of RCF

The final step in production is to release the RCF for use in LEHD infrastructure processes and LODES. The estimation and prediction model described above is applied to the entire set of residences described in Table 3. Applying the model requires replicating the steps of linking each PIK with demographic information from the ICF as well as longitudinal administrative data on place of residence.

The final RCF does not list any source information and contains the fields listed in Table 10. The RCF is unique by person, year, and residence, with person defined by a PIK, the period of residence defined as the calendar year for which the address is valid. Based on the prediction methodology, each residence is assigned both a preference weight (that sums to one by PIK) and a preference rank (starting at one, for the highest preference weight and proceeding with no ties). Each person/year/residence record includes the MAFID (where available), Census tabulation geography up to block level (where available), the length, in digits, of the geocode, and the last year in which the geocode was observed (to help with updating RCF geography to any later vintage of tabulation geography). There is an indicator for whether a residence actually was observed in the RCF address year.

4.5 Further work

Job related information from LEHD may provide further opportunities to improve this application of the person-place model. The present analysis described agreement at the person level, but it might also be calculated at the person/job level. Because administrative data are joined together from several sources, the reference date of each source may lack the coordination inherent in a household survey. The timing of jobs relative to the administrative source reference dates as well as implied commute distances between a residence and a workplace location could be used to further improve the model. For example, in the case of long distance moves, such information would favor shorter, more realistic commutes. Even if these improvements only impact a small set of records, those cases may contribute substantially to the longer, average commute distance documented for LODES relative to ACS.

[6] In practice, we link the ACS sample, by PIK to a file containing the quartely earnings histories from all jobs held in 2012 and retain only those with a match.

Table 10: RCF fields

Field	Type	Length	Label
pik	Char	9	Protected Indentification Key (PIK)
rcf_year	Num	4	Address year of RCF
addyear_obs	Num	4	Indicates observation in address year
mafid	Char	9	MAFID from best source
pref_rank	Num	8	Rank order of residential location for PIK*Year
pref_weight	Num	8	Preference weight
geocodefull	Char	15	2010 Census tabulation geography
stfidlen	Num	4	Length of geocodefull
geo_year	Num	4	Last year observed for residence geocode

5 Conclusions

The RCF methodology described in this document will be well suited for use with employer-employee matched data. The administrative data contributing to the RCF provide a high degree of coverage for the employed population and represent a range of demographics. Using the American Community Survey as training data provides a longitudinally and nationally consistent definition of residence that is strongly based on a person's regular home location. The two stage model, which is estimated for each source, includes person characteristics, and is re-estimated for each year of data, will provide highly customized predictions.

Further development of the model will allow for customization of predictions by person/job/quarter, which will provide residence predictions for each job when someone moves during a year. The LEHD program, which already makes use of multiple imputation for demographic characteristics and workplace at multi-unit employrs, will now also be able to fully represent the uncertainty of residential location in the Infrastructure Files. Public use statistics, such as LODES, will directly incorporate data from the RCF into the place of residence synthetic data model. LEHD staff will continued to develop plans for crafting these goals into a Version 1.0 for the RCF methodology.

Acknowledgments

Thanks to David Brown for sharing his previous work on residential location from administrative and other sources. Thanks to Katharine Abraham for discussion and comments at FCSM. Thanks also to the staff of CARRA for providing expertise on the CPR and MAF-ARF. Thanks to the CES data staff for preparing and providing source data for this project and to the LEHD program staff for helping to set up the development environment.

Appendix: Data Sources

5.1 Master Address File

The Geography division (GEO) at the Census Bureau maintains the Master Address File (MAF) as the frame of all residential addresses in the United States. Every year GEO releases an extract of the current MAF, consisting of a list of MAFIDs for addresses with associated characteristic information, Census tabulation geocodes, and coordinates. Because the MAF is meant to be cumulative, with addresses being added but not removed, the RCF should only need to use the most recent extract. For example, the RCF for sources up through 2013 should use the 2013 MAF extract.

5.2 American Community Survey

The ACS is the Census Bureau's continuous demographic survey, including approximately 3.5 million households annually. The sampling frame of the ACS is the MAF and the definition of residence is the home location where a household responds to the survey. Because response to the ACS is tied to occupancy, the frame of respondents is well suited to serve as a truth set for estimating a residency model based on administrative data. Respondents to the ACS provide a name, sex, and date of birth. The Census Bureau uses this information along with the residence MAFID to link respondents to a PIK. The ACS PIK rate is approximately 95 percent. The ACS microdata file also lists the response date, which makes it feasible to consider seasonal variability in a residence model.

5.3 LEHD Infrastructure Files

The Longitudinal Employer-Household Dynamics program at the Census Bureau uses job and employer information from states along with federal survey and administrative data to produce statistics on labor force dynamics including the Quarterly Workforce Indicators, LODES, and Job-to-Job Flows (Abowd et al. [2009]). States provide LEHD with quarterly files giving the earnings of all workers covered by state unemployment insurance programs. These include state and local government employment as well as approximately 96% of all private sector wage and salary employment. States also provide quarterly employer files listing establishment locations as well as industry, ownership, and size. LEHD combines these files into a Person History File, giving the earnings history of each job, and an Employer Characteristics File. LEHD also produces an Individual Characteristics file based on federal survey and administrative data that provides demographic information on workers. LEHD uses thes files to produce the public use datasets as well as for economic research.

5.4 IRS 1040

The IRS 1040 data come from the IRS 1040 income tax filings and are formatted in the source data as one record per family. Since the RCF is an individual-based, not family-based, dataset, the IRS 1040 data are transposed to create one observation for every person in an IRS 1040 household with a valid PIK. There are up to 6 PIKs per household.

The date variable is set to $(datafileyear) + 1$, since the data file is named according to the tax year, which is the previous calendar year, but the address date is as of the date of filing during the subsequent year. In this version of the RCF, only the year information is used, although additional information regarding the reference date is available in some years, notably the week of filing (the posting cycle date). This information may be used in future iterations of the RCF.

5.5 IRS 1099

The IRS 1099 data come from IRS 1099 income tax filings. Individuals that have a 1099 filing do not necessarily file a IRS 1040 tax return, thus the 1099 addresses supplement the 1040 addresses. Each 1099 filing is associated with only one individual, but an individual may have more than one filing if filing with multiple companies and/or multiple times a year. Thus, the 1099 data contains a large number of duplicates by PIK.

There are two 1099 files per year, one for the first 41 weeks of the year, the other for weeks 42 through 52. The observations are at the individual level, not the family level as in the 1040 files. The data variable is set to $(datafileyear) + 1$, since the data file is named according to the tax year, which is the previous calendar year, but the address data are as of the date of filing during the subsequent year. There is no specific filing date available in the 1099 data.

5.6 HUD PIC

The HUD PIC data come from the Department of Housing and Urban Development's Multi-Family Tenant Characteristics system. This system is used by public housing agencies to record the information from form HUD-50058, the family report. It records information on all families and the units they occupy. The HUD PIC data contain low-income individuals who may not file tax returns and thus may not be captured in the IRS 1040 or IRS 1099 data.

The input data are provided with one observation per person for every person residing in public or voucher housing. There is often more than one individual per address. The input data are delivered yearly. The reference date used is the year the file was delivered. The data contain a move-in date that could be used in future versions of the RCF to provide either an address further back in time or more longitudinally accurate address assignment.

5.7 HUD TRACS

The HUD TRACS data come from the Department of Housing and Urban Development's Tenant Rental Assistance Certification System. This is used to improve financial controls over assisted housing programs, including voucher programs. Thus, the population contained in HUD TRACS is a low income population that may not be covered in the HUD PIC data.

The input data are provided with one observation per person for every person residing in assisted housing. There are often more than one individual per address. The input data are delivered yearly. The reference date used is the year the file was delivered. The data contain a move-in date that could be used in future versions of the RCF to provide either an address further back in time or more longitudinally accurate address assignment.

5.8 HHS IHS

The Health and Human Services Indian Health Service Patient Registration provides data on the Native American population, who may not file a 1040 income tax return if they earn income exclusively on reservation land.

Observations in the IHS file are at the individual level. The reference data is the date the file was delivered, however a more detailed date is available on the file, which could be used in the future.

5.9 SSS

The Selective Service System Registration files provide data for the male population at time of registration. This may be a more comprehensive dataset of addresses of young men than the other sources, but will not capture changes in address.

5.10 HUD CHUMS

The HUD CHUMS data come from the Department of Housing and Urban Development's Computerized Home Underwriting Management System. The individuals in this input dataset are those receiving HUD mortgage assistance. Thus, they are a population of low-income individuals who are not captured in the public or voucher housing data.

The HUD CHUMS input data can contain both a borrower and a co-borrower on each observation. When there is a co-borrower, the data are reformatted to provide one person per observation with identical addresses for the borrower and co-borrower. The reference year on the HUD CHUMS data is the closing date of the loan. The input dataset contains all loans from 2000 to 2010. *Since the end data of this file is 2010, HUD CHUMS data is not used in the first iteration of the RCF.*

5.11 Medicare

The Medicare data provide information for the elderly population, who may not be captured in the other datasets if they are no longer working, and thus do not file income taxes and are not included in the other address datasets because they are not part of those populations.

The Medicare data are at the individual level. The reference date used is the date the input file was delivered. There is a residence change date in the file that may be used in future versions of the RCF to more accurately assign a date to the address.

5.12 USPS NCOA

The US Postal Service's National Change of Address (USPS NCOA) file contains records from filing change of address forms with the United States Postal Service. The most recent change of address for each individual and/or family that filed a change of address form between 2009 and 2013 is recorded in the NCOA file, with both the "to" address and the "from" address. The RCF currently uses the move effective date and the "to" address, but the "from" address can be added for later versions, creating a short, two-address panel for each observation. For the present version of the RCF, addresses are assigned to the move year if the move occurred from January to April of that year, otherwise the following calendar year is assigned to the address.

There are also some address categories within the NCOA data that are not currently used, but could provide information on the quality and/or timing of the address. These include indicators for whether an address was a temporary or permanent move and if the addresses were residential or business addresses. The address is currently only associated with a single PIK per record, but some change of address filings are for full families and others are for individuals. This information may also be used in future versions.

The street address in the NCOA file is not a single field, but 4 separate fields for the number (ncoa_input_new_prim_num), the prefix direction (ncoa_input_new_pre_dir), the street name (ncoa_input_new_prim_name), and the suffix (ncoa_input_new_suffix).

These are concatenated together to form a single street address for the output file. Some of the addresses are foreign addresses, which have a foreign address indicator. These addresses are the most frequent addresses with no assigned MAFID.

5.13 CPR/MAF-ARF

The original CPR file, produced at the Census Bureau from StARS covered the years 1999 to 2010. In 2011, LODES used a file that supplemented the 2011 MAF-ARF with the 2010 CPR. While the 2011 MAF-ARF was composed of records with a PIK and MAFID, the 2010 CPR also includes records with less precise geographic information.

References

J.M. Abowd, L. Stephens, B. Vilhuber, F. Andersson, K. McKinney, M. Roemer, and S. Woodcock. The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. Technical report, in T. Dunne, J.B. Jensen, and M.J. Roberts, eds., Producer Dynamics: New Evidence from Micro Data. Chicago: University of Chicago Press for the National Bureau of Economic Research, pp. 149-230, 2009.

J. David Brown. Synthesizing Numerous Pre-Existing Data Sources to Accurately Enumerate Nonreponding Decennial Housing Units. Technical report, United States Census Bureau, 2013.

Sonya Rastogi, Amy OHara, James Noon, Ellen A. Zapata, Cindy Espinoza, Leah B Marshall, Teresa A Schellhamer, and J. David Brown. 2010 Census Match Study. Technical report, United States Census Bureau, 2012.

Darcy Steeg Morris, Andrew Keller, and Brian Clark. An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census. Technical report, Statistical Journal of the International Association of Official Statistics, 2016, forthcoming.

D. W. Stevens. Employment that is not covered by state unemployment insurance Laws. Technical report, LEHD Technical Paper No. TP-2007-04, 2007.

Deborah Wagner and Mary Layne. The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications (CARRA) Record Linkage Software. Technical report, CARRA Working Paper Series Working Paper 2014-01, U.S. Census Bureau, 2014.