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

**by**

**Matthew R. Graham**
**U.S. Census Bureau**

**Mark J. Kutzbach**
**U.S. Census Bureau**

**Danielle H. Sandler**
**U.S. Census Bureau**

**CES 17-40 May, 2017**

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. 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. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact J. David Brown, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K034A, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov. To subscribe to the series, please click here.

**Abstract**

This paper describes the Longitudinal Employer-Household Dynamics (LEHD) program's ongoing 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 U.S. Census Bureau. 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. The discontinued file provided only a single residence per person/year, even when contributing administrative data may have contained multiple residences. This paper describes the motivation for the project, our methodology, the administrative data sources, the model estimation and validation results, and the file specifications. We find that the best prediction of the person-place model provides similar, but superior, accuracy compared with previous methods and performs well for workers in the LEHD jobs frame. We outline possibilities for further improvement in sources and modeling as well as recommendations on how to use the preference weights in downstream processing.

[*] 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 Introduction

This paper describes the Longitudinal Employer-Household Dynamics (LEHD) program's ongoing 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 U.S. Census Bureau. 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. The discontinued file provided only a single residence per person/year, even when contributing administrative data may have contained multiple residences. This paper describes the motivation for the project, our methodology, the administrative data sources, the model estimation and validation results, and the file specifications. We find that the best prediction of the person-place model provides similar, but superior, accuracy compared with previous methods and performs well for workers in the LEHD jobs frame. We outline possibilities for further improvement in sources and modeling as well as recommendations on how to use the preference weights in downstream processing.

Note that this RCFv1.1 document builds on and borrows from previous documentation, including RCFv0.5 [Graham et al., 2015] and RCFv1.0 [Graham et al., 2016].

## 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 (QWI), LEHD Origin-Destination Employment Statistics (LODES), and Job-to-Job Flows (J2J) [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 the Employment History File (EHF), which lists the earnings history of each job, and the Employer Characteristics File (ECF). LEHD also produces the 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 research into data quality and for economic analysis.[1]

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. Lastly, LODES uses the residence location of jobs to produce origin-destination tables of home-to-workplace flows, which are released using confidentiality protection measures described in Machanavajjhala et al. [2008]. These residence data are disseminated through the OnTheMap web tool, which highlights the census block level geography in LODES.[2] Comparisons of LODES with survey-based commuting statistics have considered the role that residence locations sourced from administrative records may play in differences between tabulations, warranting efforts to better understand the quality of administrative residence data [Graham et al., 2014].

From its inception, the LEHD program has used the Composite Person Record (CPR) as a source of residence data. The CPR file series, which begins in 1999, contains fields that provided a linkage between a unique person record and a place of residence. 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).[3] 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 conveyed by CARRA (in January, 2013) that were implemented by the CES Data Staff.

The 2012 process of deduplicating 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 [Graham et al., 2015]. LEHD used the output of this new process to supply the 2013-2015 cycles of LODES processing (data years 2012-2014) with residential information. Incremental research developed a new methodology, RCFv1.0 [Graham et al., 2016], which did not go into production. The iteration described here (RCFv1.1) enhances 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. LEHD first used RCFv1.1 to produce residence data for 2015 and to update residence data for 2014.

*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] Some examples of research using residence data are Andersson et al. [2014], Hellerstein et al. [2014], Hyatt [2015], Hyatt et al. [2016], and Green et al. [2017].

[2] For more information on LODES and the web tool, see `https://lehd.ces.census.gov/data/`.

# 2 Methodology

The RCF transforms a set of administrative source files listing locations for a person at a time into a file with preference weights for each person/location by reference period 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 a residence for a person and the set of sources contributing to both files omit records that could not be geocoded or assign a person identifier. The present analysis features the latest version of the person-place model, which offers more customized predictions and provides more information on the uncertainty of locations for a person than the CPR approach.

## 2.1 Baseline Rank-Order Model

The CPR limits residence data to a "best" record for each person in a year, selected by a rank-ordering or sources (developed for the StARS).[4] For processing the 2010 MAF-ARF residence data, CES used a rank 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. CES deduplicated the MAF-ARF according to the rank order and retained the highest ranking residence source for each person. In RCFv0.5, we replicate this process and implement the same rank-ordering. In the present analysis, we also produce a rank-order allocation for comparison purposes.

There are several drawbacks to LEHD processes with using the RCFv0.5 rank order method to create an annual deduplicated residence file:

- First, the availability of sources varies longitudinally and not all sources included in the Match Study are available for the RCF. New sources could not be implemented to the rank-order model without further testing. The QWI series spans 1985 to the present, so there is broad scope for longitudinal variation in source quality, which might also favor a re-ordering of sources from one year to another.
- Second, a single rank ordering 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 rank orderings.
- Third, the rank order developed in the 2010 Census Match Study is based on an April 1 reference date for a single year, but LEHD produces job statistics in QWI on a quarterly basis. Seasonal variation in the quality of the sources might alter the rank order.[5]

[3] See "Statistical Administrative Records System (October 27, 2010, 75FR 66061)" https://www.census.gov/content/dam/Census/about/about-the-bureau/SORNs/CEN-08.pdf

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

[5] The April 1 happens to be ideal for LODES, which tabulates jobs held at the beginning of the second quarter each year (with earnings in both the first and second quarter), known in LEHD as Beginning-of-Quarter employment.

- Fourth, 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, such as the assignment of establishments to workers at multi-unit employers, LEHD reflects uncertainty over tabulating statistics with weights associated with the probability of any piece of information [Abowd et al., 2009]. Furthermore, a worker may have multiple residences and jobs within a period of time, with each residence being the appropriate origin for a commute to a particular job.

## 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 an administrative records model of housing unit occupancy.[6] Brown [2013] uses a two-stage logistic model at the person-place level of location agreement, a binary outcome. The first stage estimates an equation for each administrative 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 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 an ongoing series of 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 use a residence-based frame. Because respondents must reside at an address in the mailing frame, this truth set 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.

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 cases in which 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 allow for LEHD processes to be more nuanced in the use of longitudinal information, rather than applying a series of edits.

Lastly, the current implementation is almost completely in line with the ordering of a production sequence. We limit the model estimation to a truth set combining persons in the ACS with linked administrative residences. For explanatory variables, we use LEHD job histories and demographic information already included in the LEHD Infrastructure Files, but we do not include information that is produced downstream of (and therefore dependent on) the processing of residence data.

## 2.3 Technical Description of Person-Place Model Training, Validation, and Implementation

For the population of the United States, consider the set of $N$ persons who can be assigned a unique identifier based on administrative data linkage (a requirement for a person to be linked to the LEHD job frame). These $N$ persons, indexed by $i$, may have one or more residences listed over a time period in administrative data, denoted $Admin_{ils}$, where $l$ and $s$ index the domain of locations $L$ and sources $S$, respectively. We produce the RCF for the subset labeled $N^L = \sum_{i=1}^{N} I(\sum_{l=1}^{L} \sum_{s=1}^{S} Admin_{ils})$, where the indicator function identifies persons with any listed residence from any source. We do not list a time subscript, but we assume a reference time period (e.g. year, quarter).

Because a person may have multiple administrative residences, we train a model to discriminate between residences based on person, location, and source information. As a truth set, we use residences from survey responses, denoted $Survey_{il}$. By linking the $N$ persons with survey data using unique identifiers, we define training and validation samples labeled $N^T = \sum_{i=1}^{\tau N} I(Survey_{ils}) \cdot I(\sum_{l=1}^{L} \sum_{s=1}^{S} Admin_{ils})$ and $N^V = \sum_{i=\tau N}^{N} I(Survey_{ils})$. After allocating a share $\tau$ for training and reserving a share $1-\tau$ for validation, the indicator function restricts the sets to persons with a survey-observed residence location. While $N^T$ requires that the person also have at least one administrative residence, $N^V$ makes no linkage requirement. Given that only a small share of the population is surveyed in a year, the vast majority of persons with administrative residences in $N^L$ do not appear in either $N^T$ or $N^V$, but there is some overlap.

For each person in the $N^T$ and $N^V$ we specify agreement, $\Gamma_{ils}$, 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. (Those in $N^V$ with no administrative residence are given $\Gamma_{ils} = 0$.) Agreement in Equation 1 could be specified for a range of geographic precision for $l$, including by address identifier, Census tabulation geography, county, or state. For the present analysis, we use census tabulation blocks, the most detailed geographic tabulation of residence published

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

by LEHD in LODES. As we discuss later, the time range of our residence candidates allows for a window of agreement around the reference period in recognition that most sources are only occasionally observed.

For the training sample, $N^T$, 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 $X_{ils}$, a procedure known as scoring. 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 retain characteristics that vary by person across locations, given by the vector $X_{il}$. We add indicator variables $S_{il}$ and interactions with the predicted probabilities $\hat{\Gamma}_{ils}$ for each source appearing for that person/location. We then estimate a second logistic model, for the same $N^T$ persons, specified as

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

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, $N^V$, we apply the parameter estimates from equations (2) and (3) to compute expected agreement for each person/location, where

$$\hat{\Gamma}_{il} = \frac{exp(\hat{\gamma} + \hat{\beta} X_{il} + \sum_{s=1}^{S}(\hat{\phi}_{ils} S_{il} + \hat{\lambda}_{ils} S_{il} \hat{\Gamma}_{ils}))}{1 + exp(\hat{\gamma} + \hat{\beta} X_{il} + \sum_{s=1}^{S}(\hat{\phi}_{ils} S_{il} + \hat{\lambda}_{ils} S_{il} \hat{\Gamma}_{ils}))} \tag{4}$$

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 from Equation 4 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, $\hat{W}_{il}$, specified as

$$\hat{W}_{il'} = \frac{\hat{\Gamma}_{il'}}{\sum_{l=1}^{L_i}(\hat{\Gamma}_{il})}, \tag{5}$$

where the set of weights for a person sum to unity.

We compare preference weights with two alternatives. First, we compute a modal, or "best guess" weight, defined as

$$W_{il'}^{Modal} = I(\hat{\Gamma}_{il'} = max_i(\hat{\Gamma}_{il})). \tag{6}$$

Second, we compute a rank-order weight, written as $W_{il'}^{Rank}$. For these comparison weights, a single administrative residence receives a weight of one and any others are left with a weight of zero. In Equation 6, all weight goes to the residence with the highest predicted probability. The rank-order model gives all weight to the residence with the lowest source rank-order. We resolve ties arbitrarily in either case.

In our model validation analysis, we compute an agreement rate for each weighting method across all persons in the validation sample, $N^V$, calculated as

$$\text{Agreement Rate} = \frac{1}{N^V} \sum_{i=1}^{N^V} I_{L_i > 0}(\sum_{l=1}^{L_i} \Gamma_{il} W_{il}). \tag{7}$$

Recall that $N^V$ includes all persons in that sample, even if they have no linked candidate administrative residence. The indicator function $I_{L_i > 0}$ designates those persons with no residence as contributing zero to the agreement rate.

To produce the RCF, we implement the same prediction procedures, but for the full set of persons with administrative residences, $N^L$. Specifically, we predict $\hat{\Gamma}_{il}$ for each person/residence and produce all three weights. This file, with no remaining source information, is ready for downstream processing and subsequent calculations may use the weights for computing statistics.

# 3 Data

## 3.1 Administrative Data Sources

The RCF 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, which are not available for any other LEHD process:

**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)

## 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 records identify job holders with a Protected Identification Key (PIK). Likewise, the administrative source files identify persons with a PIK. The Census Bureau assigns PIKs to administrative and survey data using personal identifying information [Wagner and Layne, 2014]. In cases where a PIK could not be assigned, data integration is not possible.

We use the Master Address File ID (MAFID), an address identifier used throughout demographic survey areas at the Census Bureau, to define a residence location. The MAF is the residence frame for both the Decennial Census and the ACS. 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 date 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 Assembling Annual Residence Files

Table 1: PIKed Input Records with a MAFID in Extract for Output Year 2015

| Year | Count (millions) |
|---|---:|
| 2008 | 2.24 |
| 2009 | 3.03 |
| 2010 | 0.66 |
| 2011 | 17.89 |
| 2012 | 42.57 |
| 2013 | 1143.59 |
| 2014 | 1221.10 |
| 2015 | 1207.86 |

In our processing system, we bring together administrative records for our sources in a range of data series and file formats. Some files include a series that spans multiple years, while others include multiple series. A given source may be composed of multiple input series that arrive in different file formats from year-to-year.[7] In Table 1, we first tabulate the number of address records received for all sources across all years to give an indication of the current state of longitudinal coverage. The extract is for files with a reference date from 2013 onwards, with some earlier coverage only from series that extend back prior to 2013. Coverage must be at least up to the final year of RCF being produced, which is 2015 in this case. All subsequent documentation will only be for the years necessary to produce the 2015 RCF and to update the 2014 RCF, which includes 2013 to 2015. Input data exist outside the range shown in Table 1, however they are outside the scope of the current project.

[7] Almost every source has a different structure. NCOA spans multiple years and includes two information series (move-ins and move-outs). The IRS 1040 series for a year consists of multiple files, each covering different spans of weeks. The HUD files switched formats from separate files to a joint file.

Table 2: Record Count by Source with MAFID Merge Rate in 2013

| Source | Count (millions) | Merge Rate to MAF |
|---|---|---|
| Medicare | 56.9 | 0.888 |
| NCOA | 39.6 | 0.893 |
| IHS | 6.6 | 0.732 |
| IRS1040 | 288.6 | 0.913 |
| IRS1099 | 724.5 | 0.874 |
| HUDPIC | 8.1 | 0.850 |
| SSS | 16.6 | 0.905 |
| HUDTRACS | 2.5 | 0.896 |
| Total | 1143.6 | 0.885 |

Table 3: Record Count by Source with MAFID Merge Rate in 2014

| Source | Count (millions) | Merge Rate to MAF |
|---|---|---|
| Medicare | 59.1 | 0.864 |
| NCOA | 38.2 | 0.903 |
| IHS | 7.0 | 0.568 |
| IRS1040 | 289.9 | 0.903 |
| IRS1099 | 799.5 | 0.865 |
| HUDPIC | 8.1 | 0.843 |
| SSS | 16.8 | 0.897 |
| HUDTRACS | 2.6 | 0.904 |
| Total | 1221.1 | 0.874 |

Table 4: Record Count by Source with MAFID Merge Rate in 2015

| Source | Count (millions) | Merge Rate to MAF |
|---|---|---|
| Medicare | 60.4 | 0.871 |
| NCOA | 19.2 | 0.907 |
| IHS | 7.3 | 0.565 |
| IRS1040 | 290.8 | 0.913 |
| IRS1099 | 802.6 | 0.868 |
| HUDPIC | 7.9 | 0.893 |
| SSS | 16.8 | 0.905 |
| HUDTRACS | 3.0 | 0.606 |
| Total | 1207.9 | 0.878 |

After designating the records by source and year, we are ready to begin assembling our set of residence candidates. Tables 2, 3, and 4 list the count of records with a PIK retained from each source for the years 2013, 2014, and 2015, respectively. When reading in the source files, we only retained records with a PIK. Of those, about 88 percent overall had a MAFID, which we use to link with 2010 Tabulation Geography (e.g. state, county, census tract, census block). Over 99% of these records have residence information precise enough to be geocoded to a census block.[8] The tables provide the match rate for each source. Some sources were notable in having a lower percentage of records with valid MAFIDs. In 2015, these include IHS (56.5 percent) and HUD TRACS (60.1 percent).

Table 5: Residences in 2013: Unique Records by PIK/Source/MAFID

| Unique Identifier | Count (millions) |
|---|---|
| PIK | 298.6 |
| PIK/source | 560.3 |
| PIK/source/MAFID | 620.4 |

[8]For the present analysis, we retain this small set of records merged to the MAF that still have a missing or incomplete geocode and code them as lacking agreement with the survey location.

Table 6: Residences in 2014: Unique Records by PIK/Source/MAFID

| Unique Identifier | Count (millions) |
|---|---|
| PIK | 298.6 |
| PIK/source | 559.4 |
| PIK/source/MAFID | 620.6 |

Table 7: Residences in 2015: Unique Records by PIK/Source/MAFID

| Unique Identifier | Count (millions) |
|---|---|
| PIK | 302.0 |
| PIK/source | 559.1 |
| PIK/source/MAFID | 613.6 |

We produce residence frames for 2013, 2014, and 2015 consisting of the set of unique source/address records per PIK, where each residence has a valid MAFID. We consolidate duplicate MAFIDs for the same PIK from the same source, which we consider to be redundant. Tables 5, 6, and 7 list the record count for the 2013, 2014, and 2015 files by the count of unique PIKs, unique instances of a PIK having a source (PIK/source), and unique residences for a PIK/source (PIK/source/MAFID). In 2015, the inflation of 302 million PIKs to 559 million PIK/sources and 614 million PIK/source/MAFIDs gives a sense of the depth of coverage across sources.[9] Compared to the larger total in the record count for Table 4, it also gives a sense of the degree of duplication within a PIK/source/MAFID.[10]

## 3.4 Assembling Candidate Sets for Reference Years

We introduce a concept of a "reference" period for the RCF that designates the year for which residences are considered to be candidates. For example, LODES production for 2015 jobs would use residences with a 2015 reference year. To allow for the occasional nature of residence observations and the potential misalignment of pay periods (used for reporting) and calendar years in which jobs were held, we broaden the candidate set to include residences observed prior to or after the reference year.[11]

Table 8: Reference Year 2014: Unique Records by PIK/MAFID/Source

| Unique Identifier | Count (millions) |
|---|---|
| PIK | 327.1 |
| PIK/MAFID | 555.7 |
| PIK/MAFID/source | 923.7 |

Table 9: Reference Year 2015: Unique Records by PIK/MAFID/Source

| Unique Identifier | Count (millions) |
|---|---|
| PIK | 315.0 |
| PIK/MAFID | 468.3 |
| PIK/MAFID/source | 763.7 |

We assemble our candidate frame of residences for a set of persons, $N^L$, by combining up to three adjacent years longitudinally. Tables 8 and 9 give the count of unique records by PIK, PIK/MAFID, and PIK/MAFID/Source in the candidate sets for reference years 2014 (years 2013-2015) and 2015 (years 2014-2015), respectively. In each table, the count of PIKs is $N^L$, the count of PIK/MAFIDs is the size of the candidate set for scoring the second stage of the person-place model (see Equation 3), and the count of PIK/MAFID/Sources is size the candidate set for scoring the first stage of the person-place model (see Equation 2). For example, we combine the 614 million records in the 2015 file described above with a corresponding file for 2014 (there was no 2016 file

[9] These totals are in line with the 2010 Census Match Study [Rastogi et al., 2012], which found 302 million records with a PIK and MAFID using both federal and commercial source data.

[10] This point in processing would be the stage where CPR or MAF-ARF processing would deduplicate the PIKs within a residence in a year, using the rank ordering to select a single residence. The resulting record count for 2015 would match the first row of Table 7.

[11] One future enhancement might be to narrow the this time period to a reference quarter, the most detailed time scale for LEHD. This could include using more detailed information on dates from the input files. Another approach might be to specify a reference period including all quarters for which a job was active. In that case, candidate residences would include all reported locations for the duration of the job as well some earlier and later reports.

available yet). The combined file has 764 million unique PIK/source/MAFID combinations, with 468 million unique PIK/MAFIDs and 315 million unique PIKs. Likewise we assemble a three-year file around 2014, with a 2013 to 2015 window, that has 924 million unique records.

Table 10: Candidate Set at PIK/MAFID/Source Level for Reference Year 2014

| Source | 3-Year Count (millions) |
|---|---|
| Medicare | 65.9 |
| NCOA | 78.0 |
| IHS | 4.6 |
| IRS1040 | 367.3 |
| IRS1099 | 373.4 |
| HUDPIC | 11.5 |
| SSS | 19.5 |
| HUDTRACS | 3.5 |
| Total | 923.7 |

Table 11: Candidate Set at PIK/MAFID/Source Level for Reference Year 2015

| Source | 3-Year Count (millions) |
|---|---|
| Medicare | 58.0 |
| NCOA | 48.4 |
| IHS | 3.3 |
| IRS1040 | 308.2 |
| IRS1099 | 316.1 |
| HUDPIC | 9.8 |
| SSS | 17.1 |
| HUDTRACS | 2.9 |
| Total | 763.7 |

Tables 10 and 11 give the record totals by source for the longitudinally aggregated files for reference years 2014 and 2015, respectively. The record count for each source is the candidate set for each source-specific version of the first stage of the person-place model.

## 3.5 Defining Residence Characteristics

In combining the longitudinal information, we retain the residence pattern across years as a three-digit "triad" that will serve as an input to the estimation model. Our triads may also be described as a 3-tuple, defined for reference year $t$ as indicators for whether a residence is observed in the adjacent years, written as $(t-1, t, t+1)$. For example, in the 2014 file, a person with the same MAFID from the IRS 1040 in both 2013 and 2014, but not in 2015, would have the value "110" for that record, where ones indicate that the MAFID was reported by that source in a year. In our estimation model, we include these triads as explanatory variables.

Because our person-place model makes use of person characteristics, we supplement the PIK/source/MAFID level file with demographic information from the National Individual Characteristics File (NICF) at the PIK level. LEHD produces the NICF, which provides place of birth (country), date of birth, sex, race, ethnicity, and educational attainment, from NUMIDENT and demographic files (currently the 2000 Census). The NICF is considered to be invariant longitudinally (though it is reprocessed as new information becomes available) so we do not need to link the PIKs with information that varies by year. We calculate age on April 1 of the reference year based on date of birth (to align with timing for LODES, the most detailed user of residence data). We produce categorical variables from each demographic variable for use in our models (we do not use educational attainment because the imputation rate is especially high).

Note that for production sequencing, we use the NICF because the ICF itself is downstream of the residence data.[12] As such, our model is limited to using observed demographic information from NICF and does not have access to the imputed values from the ICF. We set any variable with no observed value to a missing value, "X", and use that as the omitted value in our estimation model. Approximately two-thirds of persons (and three-quarters of workers) in our residence files link to an observed value for all of the demographic variables we use. Having an observed value is most common for age, sex, and place of birth.

[12] In practice, we currently use the ICF where the "imputed" indicator is set to "1", indicating that the record was observed, and supplement the file with the ICF nonworkers file and only where the "flag" is set to "0". If a person is in both, we prefer the ICF variables to the ICF nonworkers variables. Future versions of RCF should make use of a new demographic input file that is currently under development.

Lastly, we link our person records with LEHD earnings records, by PIK, to create an indicator for whether a person had employment in a given year. Because the earnings records themselves are not downstream of the residence data, this linkage is not circular and the information should be available at the time of processing.[13] Future work could develop more informative variables relating to work history that might help with distinguishing between residences in the case of long distance moves.

## 3.6 Analysis of Residence Candidates

Table 12: Count of MAFIDs per PIK for Reference Year 2014, by Same and Three-Year Range

| Count of MAFIDs for a PIK | 1-Year Count (mil.) | 3-Year Count (mil.) | 1-Year Share (%) | 3-Year Share (%) |
|---|---|---|---|---|
| 0 | 28.5 | . | 8.7 | . |
| 1 | 228.6 | 193.2 | 69.9 | 59.1 |
| 2 | 52.2 | 78.5 | 16.0 | 24.0 |
| 3 | 13.6 | 32.2 | 4.1 | 9.9 |
| 4 | 3.3 | 13.5 | 1.0 | 4.1 |
| 5+ | 0.9 | 9.7 | 0.3 | 3.0 |
| Total | 327.1 | 327.1 | 100.0 | 100.0 |

Table 13: Count of MAFIDs per PIK for Reference Year 2015, by Same and Three-Year Range

| Count of MAFIDs for a PIK | 1-Year Count (mil.) | 3-Year Count (mil.) | 1-Year Share (%) | 3-Year Share (%) |
|---|---|---|---|---|
| 0 | 13.0 | . | 4.1 | . |
| 1 | 236.4 | 212.6 | 75.0 | 67.5 |
| 2 | 51.1 | 68.4 | 16.2 | 21.7 |
| 3 | 11.6 | 22.4 | 3.7 | 7.1 |
| 4 | 2.3 | 7.9 | 0.7 | 2.5 |
| 5+ | 0.5 | 3.7 | 0.2 | 1.2 |
| Total | 315.0 | 315.0 | 100.0 | 100.0 |

Beyond assembling and editing the input sources, the main added value of the RCF is in discriminating between multiple MAFIDs for the same person in the same time period. To give a sense of the scope for disagreement, Tables 12 and 13 list the distribution of persons by count of different residences across all sources for reference years 2014 and 2015. The 2014 table is for the 327.1 million persons with a residence in 2013, 2014, or 2015, while the 2015 table includes only the 315.0 million with a residence in 2014 or 2015 (2016 was not yet available). The first column gives the count, from zero up to 5 or more, the next two columns give frequencies (in millions) and the last two give percentages.

Within each summary set, the left column considers only MAFIDs appearing in the reference year, while the right column considers those in both the reference year and in adjacent years. Because the set of persons must have a residence in one of the relevant years, the right columns (by count or percent) include no persons with zero MAFIDs, while the left cases do have persons with no MAFIDs. By including multiple years, we augment the reference year MAFIDs with PIKs that did not have a residence in the reference year (for 4.1 percent of PIKs in 2015) and also reduce the share of persons with only one MAFID.

For Tables 12 and 13, within the reference year (1-Year Share), 69.9 and 75.0 percent of persons, respectively, have only one residence. Augmenting the file with adjoining years (3-Year Share) adds more residences and reduces the abundance of these one-residence persons to 59.1 and 67.5 percent of the sample. The remaining individuals, with multiple MAFIDs, are the motivation for our modeling approach described below. Note that some of the discrepancies in residences within a person here are due to the precision of MAFIDs. When residence is defined by Census block, rates of having only one residence in 2014 and 2015 rise to 73.0 and 78.6 percent for the 1-Year frame and 62.7 and 71.3 percent for the 3-Year frame.

[13] In the present production process, we use the Person History Enhanced Across SEIN and Non-SEIN Transitions (PHEASANT) file to consolidate state level Person History Files (PHF_B) and retain quarterly total earnings across all jobs held in the reference year. Technically, the PHF_B is downstream of RCF processing, so it would be ideal to use the EHF. The PHEASANT is designed around the PHF_B and simplifies assembly of the earnings records, but ideally a similar assembly process would be adapted for the EHF.

Table 14: Longitudinal Residence Histories (Triads), by Source in Reference Year 2014

| sourcecode(Source) | Table of sourcecode by src_year_hist<br>src_year_hist(Year History Pattern of MAFID for Source and PIK) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 001 | 010 | 011 | 100 | 101 | 110 | 111 | Total |
| Medicare | 10.40 | 1.93 | 11.00 | 12.05 | 0.10 | 6.27 | 58.25 | |
| NCOA | 18.48 | 37.33 | 1.27 | 37.97 | 1.48 | 3.13 | 0.35 | |
| IHS | 5.85 | 1.39 | 12.51 | 27.44 | 0.11 | 2.16 | 50.54 | |
| IRS1040 | 13.22 | 6.80 | 10.34 | 16.09 | 1.61 | 7.09 | 44.86 | |
| IRS1099 | 14.95 | 6.28 | 11.01 | 15.34 | 1.16 | 8.67 | 42.58 | |
| HUDPIC | 24.42 | 6.49 | 9.24 | 15.00 | 1.41 | 17.88 | 25.57 | |
| SSS | 11.28 | 0.57 | 12.08 | 12.31 | 0.03 | 10.02 | 53.70 | |
| HUDTRACS | 14.88 | 13.26 | 8.11 | 18.49 | 0.54 | 24.01 | 20.70 | |
| Total | 1.315E8 | 8.019E7 | 9.16E7 | 1.599E8 | 1.163E7 | 6.996E7 | 3.789E8 | 9.237E8 |

Table 14 shows the longitudinal history pattern of addresses by source file. The header gives the three-digit triads that indicate whether the address was present in 2013, 2014, and 2015. So, if an address is present in 2013 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 12.05 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. The total row gives person counts by each triad type.

Table 15: Longitudinal Residence Histories (Triads), by Source in Reference Year 2015

| sourcecode(Source) | Table of sourcecode by src_year_hist<br>src_year_hist(Year History Pattern of MAFID for Source and PIK) | | | |
|---|---|---|---|---|
| | 010 | 100 | 110 | Total |
| Medicare | 11.94 | 9.32 | 78.74 | |
| NCOA | 32.17 | 65.22 | 2.61 | |
| IHS | 8.21 | 4.89 | 86.89 | |
| IRS1040 | 17.67 | 16.55 | 65.78 | |
| IRS1099 | 19.03 | 17.67 | 63.30 | |
| HUDPIC | 30.39 | 28.67 | 40.95 | |
| SSS | 12.90 | 12.09 | 75.02 | |
| HUDTRACS | 18.92 | 45.73 | 35.35 | |
| Total | 1.431E8 | 1.502E8 | 4.705E8 | 7.637E8 |

Table 15 repeats the tabulation of the previous table, but only for PIK/sources with a MAFID appearing in 2014 or 2015. As a result, the triads are limited to those with a residence in $t-1$ and/or $t$, relative to the reference year.

## 3.7 Preparing the Training and Validation Samples

To train the predictive residence model, we estimate models for administrative record agreement with a linked person's residence in the ACS. Because only the 2014 ACS files were available in time for production (see below), our model development, training, and validation, are limited to the 2014 reference year. While $N^T$ and $N^V$ (see Section 2.3) are both for the 2014 reference year, we will use these models to score both the 2014 and 2015 candidate files to produce an RCF for each year. The model specifications will differ according to the preceeding and succeeding years of administrative data available for scoring each reference year.

We draw a person frame for our truth set of residences from 5.5 million respondents to the ACS in 2014 (ACS responses for the reference year were not yet available).[14] Because we are focused on residence prediction for an administrative jobs frame, we first limit the sample to the 5.0 million respondents who can be linked to a PIK using personal identifying information. The PIK rate for the ACS in 2014 is 91.4%. The only information we retain from the ACS is the quarter of response within 2014 and the edited place of residence Census block from the tabulation geography in ACS. We randomly partition this set of 5 million PIKs and their associated 13.6 million PIK/source/MAFID records into a 70 percent training sample and 30 percent validation sample.

[14] As the ACS is year-round, we limit the extract to only those in the 2014 ACS actually responding with a reference date in 2014 (99 percent of persons).

We retain PIKs in the ACS sample even if those persons do not link to any place of residence in our candidate frame of PIK/source/MAFID records for 2013 to 2015. Given the high likelihood of a PIKed record in ACS having some administrative residence data, less than one percent of PIKs fail to match. These persons do not contribute to training, but we do include them in the validation sample to give a complete assessment of the quality of our matching.

As the dependent variable of our training model and the focus for validation, we construct an indicator for whether a particular administrative MAFID corresponds with the geography of an ACS residence. We use census block level precision for our measure of agreement, so if a MAFID falls in the same block as an ACS response, we consider the candidate pair to be a match. Although the ACS includes demographic and employment information, we only use the variables linked to the candidate set from the NICF and LEHD, as described above.[15]

### 3.8 Assembly of Input Data

The RCF processes use a wide range of federal administrative and survey data source files on residence location and person characteristics that must be assembled on a single production environment. Production for a 2015 reference year did not begin until September of 2016, when the complete residence files for 2015 were available (following pre-processing by CARRA and CES). Other administrative data required for production include the LEHD earnings records and worker characteristics. The PHF_B, or alternately the EHF, are typically available with a six month lag, so earnings records for all of 2015 for most states were available in time for processing. The observed characteristics in the ICF are updated less frequently, so using the latest extract from ICF is less time-sensitive. Ideally, we would do our ACS-based training and validation for responses in the same year as our RCF frame. However, the processing of administrative records and ACS files is not synchronized and the ACS is often not available by when LEHD production needs the latest RCF. In recent years, the ACS along with the necessary PIK crosswalk has not been available until September of the year after the planned reference year.

## 4 Training, Validation, and Implementation

### 4.1 Specifications

We execute both the baseline *rank-order* methodology and the *person-place* methodology described in Section 2. In Graham et al. [2016], we considered a range of model specifications and evaluated their relative quality by comparing results for the validation sample. Relative to the simple rank-order model, we considered the importance of adding longitudinal residences and the value of including the residence history triads (e.g. "110", "101") in the first and second stages of the model (see Equations 2 and 3). The largest gains in accuracy of predicting the ACS residence came from including longitudinal information and from including the triads in the first stage. The source-specific triads in the second stage only added marginal accuracy at the cost of dozens of additional parameters, many of which were not precisely estimated.

For our present implementation, we retain the longitudinal residences and triads in the first stage, but consolidate the triads in the second stage across sources. The last specification in Graham et al. [2016] (see column 5, Table 8) included a variable for each source/triad combination (as well as an interaction with predicted probability). We consolidate the triads so there is only one triad present for a PIK/MAFID across all sources. We create a vector of indicators that may vary by person across locations included as $X_{il}$ in Equation 3. These triads are in addition to the source and source-by-probability variables, or $S_{il}$ and $S_{il}\hat{\Gamma}_{ils}$. We find that this consolidated-triad model performs just as well as the most complex model in Graham et al. [2016], so we only present results for this new approach here.

Below, we present parameter estimates for the consolidated-triad person-place model fit on the 70 percent training sample (the survey sample $N^T$) for reference year 2014. This model explains Census block agreement with our ACS extract from 2014 using MAFIDs from 2013 to 2015. We label this three-year specification as "111", indicating that it includes longitudinal information from both the reference year and both adjacent years. We also estimate a two-year specification that we label as "110", which only uses information from the previous and reference years, or 2013 and 2014. We will use the "111" specification to score the 2014 RCF and the "110" specification to score teh 2015 RCF. We present only the "111" version here because it provides a more comprehensive illustration of coefficients representing longitudinal information.

[15] We do not currently make any age restrictions for the candidate set, though further robustness checks might investigate whether the model should be estimated separately by age group or other characteristics.

## 4.2 Estimation

Table 16: First Stage Person-Place Estimates for IRS 1040 Source with Model 111

| Characteristic | Variable | Ratio w.r.t. Omitted | Log odds ratio | 95% C.I. lower bound | 95% C.I. upper bound |
|---|---|---|---|---|---|
| Triad (source/MAFID) 001 | src_year_hist | 001 vs 111 | 0.080 | 0.080 | 0.081 |
| Triad (source/MAFID) 010 | src_year_hist | 010 vs 111 | 0.095 | 0.094 | 0.096 |
| Triad (source/MAFID) 011 | src_year_hist | 011 vs 111 | 0.305 | 0.302 | 0.308 |
| Triad (source/MAFID) 100 | src_year_hist | 100 vs 111 | 0.027 | 0.027 | 0.028 |
| Triad (source/MAFID) 101 | src_year_hist | 101 vs 111 | 0.280 | 0.275 | 0.286 |
| Triad (source/MAFID) 110 | src_year_hist | 110 vs 111 | 0.180 | 0.179 | 0.182 |
| Age <14 | age_c | 0 vs X | 0.955 | 0.551 | 1.657 |
| Age 14 to 18 | age_c | 1 vs X | 0.818 | 0.472 | 1.418 |
| Age 18 to 24 | age_c | 2 vs X | 0.421 | 0.243 | 0.731 |
| Age 25 to 34 | age_c | 3 vs X | 0.909 | 0.524 | 1.577 |
| Age 35 to 44 | age_c | 4 vs X | 1.228 | 0.708 | 2.129 |
| Age 45 to 54 | age_c | 5 vs X | 1.510 | 0.871 | 2.619 |
| Age 55 to 64 | age_c | 6 vs X | 1.589 | 0.917 | 2.756 |
| Age 65 to 74 | age_c | 7 vs X | 1.577 | 0.909 | 2.735 |
| Age >74 | age_c | 8 vs X | 1.247 | 0.719 | 2.162 |
| Female | sex | F vs X | 0.934 | 0.550 | 1.585 |
| Male | sex | M vs X | 0.930 | 0.548 | 1.579 |
| White | race | 1 vs X | 0.987 | 0.965 | 1.010 |
| Black | race | 2 vs X | 0.941 | 0.918 | 0.965 |
| AIAN | race | 3 vs X | 0.707 | 0.681 | 0.735 |
| Asian | race | 4 vs X | 0.939 | 0.911 | 0.969 |
| NHPI | race | 5 vs X | 0.936 | 0.839 | 1.045 |
| Two or more | race | 7 vs X | 0.959 | 0.925 | 0.995 |
| Hispanic | ethnicity | H vs X | 1.134 | 1.110 | 1.159 |
| non-Hispanic | ethnicity | N vs X | 1.087 | 1.062 | 1.114 |
| Foreign Born | native | F vs X | 1.083 | 0.813 | 1.442 |
| Native Born | native | N vs X | 0.986 | 0.741 | 1.313 |
| Employed in LEHD | emprefyr | 1 vs 0 | 0.989 | 0.982 | 0.996 |

Table 16 gives the log odds ratios for the first stage logistic regression of the IRS 1040 model, as specified by Equation 2. The dependent variable is agreement of a 1040 candidate with ACS residence at the PIK/MAFID/source level. Though we estimate the model separately for each of our sources, we only present results from one source here. We regard these as descriptive rather than causal estimates, and present them to provide insights into the model structure rather than as evidence of any particular feature in the data. For each explanatory variable, we describe the characteristic, name the variable as coded, state the comparison of a categorical value versus an omitted value (for categorical variables only), and provide the log odds ratio along with a 95% Confidence Interval. For many categorical variables linked from the NICF, the omitted category is "X," which, as noted above, indicates that no value was observed in administrative data linked to that PIK.

Reviewing the log odds ratios, all of the listed longitudinal triad estimates are less than one, indicating that the best pattern is the omitted category, "111", a residence present in all three years. The source is more accurate for middle aged respondents than younger respondents, with those aged 18 to 24 having the worst match rate (likely because of mobility out of households where they are a dependent). Other groups with noticeably lower predicted accuracy are American Indian and Alaskan Natives and those with unknown ethnicity. While these results are descriptive of the IRS 1040 source, the same parameters may have different predictions for accuracy in other sources.

Table 17 gives the log odds ratios for the second stage logistic regression, as specified by Equation 3. The dependent variable is agreement of a candidate with ACS residence at the PIK/MAFID level. In addition to the longitudinal triads consolidated across sources, the model includes indicator variables for each of the source files and interactions with the probability predictions 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. The relative sizes of the main and interaction effects give an indication of the relative importance of simply having a source versus the importance of the individual information associated with the accuracy of the source. For example, with IHS and SSS, the interaction is substantially more important than the source itself. With SSS, for example, this may be because registration is current mainly for young men. In contrast, NCOA and the IRS 1099 are more broadly of high quality.

Table 17: Second Stage Person-Place Estimates of Model 111

| Characteristic | Variable | Ratio w.r.t. Omitted | Log odds ratio | 95% C.I. lower bound | 95% C.I. upper bound |
|---|---|---|---|---|---|
| Triad (MAFID) 001 | maf_year_hist | 001 vs 111 | 0.617 | 0.612 | 0.623 |
| Triad (MAFID) 010 | maf_year_hist | 010 vs 111 | 0.521 | 0.515 | 0.528 |
| Triad (MAFID) 011 | maf_year_hist | 011 vs 111 | 0.760 | 0.754 | 0.766 |
| Triad (MAFID) 100 | maf_year_hist | 100 vs 111 | 0.273 | 0.269 | 0.276 |
| Triad (MAFID) 101 | maf_year_hist | 101 vs 111 | 0.823 | 0.808 | 0.838 |
| Triad (MAFID) 110 | maf_year_hist | 110 vs 111 | 0.538 | 0.532 | 0.544 |
| Medicare | source_2 | n.a. | 0.362 | 0.353 | 0.370 |
| Medicare*Prob | source_2_p | n.a. | 8.169 | 7.918 | 8.428 |
| NCOA | source_3 | n.a. | 0.528 | 0.510 | 0.547 |
| NCOA*Prob | source_3_p | n.a. | 1.331 | 1.231 | 1.441 |
| IHS | source_4 | n.a. | 0.123 | 0.116 | 0.132 |
| IHS*Prob | source_4_p | n.a. | 30.800 | 26.382 | 35.957 |
| IRS1040 | source_5 | n.a. | 0.440 | 0.435 | 0.444 |
| IRS1040*Prob | source_5_p | n.a. | 29.145 | 28.732 | 29.564 |
| IRS1099 | source_6 | n.a. | 0.529 | 0.523 | 0.534 |
| IRS1099*Prob | source_6_p | n.a. | 5.423 | 5.338 | 5.510 |
| HUDPIC | source_8 | n.a. | 0.350 | 0.330 | 0.372 |
| HUDPIC*Prob | source_8_p | n.a. | 32.316 | 29.689 | 35.175 |
| SSS | source_10 | n.a. | 0.109 | 0.105 | 0.113 |
| SSS*Prob | source_10_p | n.a. | 20.535 | 18.966 | 22.233 |
| HUDTRACS | source_11 | n.a. | 0.488 | 0.449 | 0.532 |
| HUDTRACS*Prob | source_11_p | n.a. | 18.309 | 15.983 | 20.975 |

## 4.3 Validation

Having estimated the "110" and "111" specifications of the person-place model on the 70 percent training sample, we now apply the model to make predictions for the 30 percent validation sample, also for the 2014 reference year. Section 2.3 describes our validation methodology and the construction of weights. For the validation sample (the survey sample $N^V$), we use the agreement rate defined in Equation 7 to assess prediction quality. The agreement rate for any weighting method falls between zero and one, with rates closer to one indicating better agreement. Equations 6 and 5 define the modal and preference weights, respectively. We contrast these with a weighting based on the rank-order model, as a baseline.[16]

Table 18: Model 110 Agreement Rate Versus Alternatives for ACS Validation Sample, All Persons

| Measure of Agreement with ACS Residence | MEAN | STD | MIN | MAX | Obs. (millions) |
|---|---|---|---|---|---|
| Person−place model, modal weight | 0.782 | 0.413 | 0 | 1 | 1.5 |
| Person−place model, preference weight | 0.749 | 0.386 | 0 | 1 | 1.5 |
| Rank−order weight | 0.767 | 0.422 | 0 | 1 | 1.5 |
| Indicator for MAF in any year | 0.947 | 0.225 | 0 | 1 | 1.5 |
| Indicator for MAF in reference year | 0.913 | 0.282 | 0 | 1 | 1.5 |
| Perfect weighting | 0.824 | 0.381 | 0 | 1 | 1.5 |
| Random weighting | 0.680 | 0.397 | 0 | 1 | 1.5 |

We first present validation results for the "110" specification. For the full ACS validation sample (with no employment restriction), Table 18 gives the agreement rates for each weighting method as well as several benchmarking measures. The first three rows give agreement for the modal, preference, and rank-order weights, respectively.

We compare these weights with several bounding measures below. To illustrate the extent of blocking to the candidate set, we provide the share of persons with a MAFID over the any of the three years (2013-2015) and a MAFID in the reference year. As an upper bound for the agreement rate, we compute a perfect agreement rate as $W_{il}^{Perfect} = \Gamma_{il}$ (so that the weight is one for an agreeing MAFID and zero otherwise). We also report an uninformed or "random" agreement rate, placing equal weight on each MAFID with $W_{il}^{Rand} = 1/L_i$. The agreement rate for all of our three model weights should fall between the last two measures.

[16]Note that although the rank-order source list was provided by CARRA, these estimates are not produced by CARRA and the results are not to be interpreted as an evaluation of a methodology that was developed for other purposes. Rather, we include the rank-order model as a baseline from RCFv0.5, providing a baseline had there been no further model development work.

For the "110" model specification, our person-place weights are similar to the rank-order weights, with the modal weights performing moderately better.[17] We find an agreement rate of 0.782 for modal weights, meaning that we would expect 78.2 percent of "best guess" residence predictions to match the ACS location. This rate compares to 0.749 for the preference weights and 0.767 for the rank-order weights. These measures are substantially above random matching, at 0.680, but have room to improve relative to perfect matching, at 0.824.

To put agreement rates relative to these benchmarks in perspective, consider the following achievement index:

$$\text{Achievement Index} = 100 * \frac{\text{Agreement Rate} - \text{Random Matching Rate}}{\text{Perfect Matching Rate} - \text{Random Matching Rate}}, \tag{8}$$

which gives the percent of perfect matching explained relative to random guessing (on a 0 to 100 percent scale). By this Achievement Index, the modal weight method achieves 71 percent of its potential, with $0.708 = (0.782 - 0.680)/(0.824 - 0.680)$. Likewise, the preference weight and rank-order weight methods achieve 48 and 60 percent, respectively.

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 if the set of sources were updated.

The relatively worse performance of the preference 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. A further possibility is that some low probability MAFIDs may be sufficiently unlikely that they should be omitted from normalization and given a weight of zero. Experimentation could be used to find an optimal cutoff for omitting low probability records. After calculating predicted probabilities with Equation 4, low scoring records would be removed (or set to zero) before calculating preference weights with Equation 5).

Table 19: Model 110 Agreement Rate Versus Alternatives for ACS Validation Sample, Employed Persons

| Measure of Agreement with ACS Residence | MEAN | STD | MIN | MAX | Obs. (millions) |
|---|---|---|---|---|---|
| Person−place model, modal weight | 0.823 | 0.382 | 0 | 1 | 0.7 |
| Person−place model, preference weight | 0.776 | 0.352 | 0 | 1 | 0.7 |
| Rank−order weight | 0.812 | 0.391 | 0 | 1 | 0.7 |
| Indicator for MAF in any year | 0.987 | 0.113 | 0 | 1 | 0.7 |
| Indicator for MAF in reference year | 0.965 | 0.183 | 0 | 1 | 0.7 |
| Perfect weighting | 0.872 | 0.334 | 0 | 1 | 0.7 |
| Random weighting | 0.688 | 0.375 | 0 | 1 | 0.7 |

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 earnings from at least one LEHD job in the reference year. We present the results for employed respondents in Table 19. For this worker sample, we find the same pattern as before, but with higher agreement rates. The agreement rate for modal weights is 0.823, with an Achievement Index of 73 percent.[18]

Table 20: Model 111 Agreement Rate Versus Alternatives for ACS Validation Sample, All Persons

| Measure of Agreement with ACS Residence | MEAN | STD | MIN | MAX | Obs. (millions) |
|---|---|---|---|---|---|
| Person−place model, modal weight | 0.801 | 0.399 | 0 | 1 | 1.5 |
| Person−place model, preference weight | 0.759 | 0.360 | 0 | 1 | 1.5 |
| Rank−order weight | 0.767 | 0.422 | 0 | 1 | 1.5 |
| Indicator for MAF in any year | 0.966 | 0.181 | 0 | 1 | 1.5 |
| Indicator for MAF in reference year | 0.913 | 0.282 | 0 | 1 | 1.5 |
| Perfect weighting | 0.867 | 0.339 | 0 | 1 | 1.5 |
| Random weighting | 0.711 | 0.374 | 0 | 1 | 1.5 |

[17] Note that they agreement rates reported here are not directly comparable to those in Graham et al. [2016], as those results were for different years and also considered an ACS sample that was restricted by age.

[18] The Acheivement Index for the modal, preference, and rank-order models are 73.4, 47.8, and 67.4 for Table 19.

Table 21: Model 111 Agreement Rate Versus Alternatives for ACS Validation Sample, Employed Persons

| Measure of Agreement with ACS Residence | MEAN | STD | MIN | MAX | Obs. (millions) |
|---|---|---|---|---|---|
| Person−place model, modal weight | 0.838 | 0.369 | 0 | 1 | 0.7 |
| Person−place model, preference weight | 0.783 | 0.323 | 0 | 1 | 0.7 |
| Rank−order weight | 0.812 | 0.391 | 0 | 1 | 0.7 |
| Indicator for MAF in any year | 1.000 | 0.000 | 1 | 1 | 0.7 |
| Indicator for MAF in reference year | 0.965 | 0.183 | 0 | 1 | 0.7 |
| Perfect weighting | 0.917 | 0.276 | 0 | 1 | 0.7 |
| Random weighting | 0.719 | 0.349 | 0 | 1 | 0.7 |

Validation results for the three-year window, or "111" specification, again for the 2014 reference year, are presented in Table 20, for all persons, and Table 21, for workers. While not suitable for producing the latest year, this specification could update an earlier production run that did not have the third year of residence data. Indeed, the person-place model with modal weights for the "111" specification has higher agreement rates for the all persons and workers samples of 0.801 and 0.838, respectively (compared with 0.782 and 0.823 for the "110" specification). The person-place models with the "111" specification perform even better relative to the rank-order model in terms of the Achievement Index.[19] The improvement of the "111" over the "110" specification confirms that, once all three years are available, replacing the initial RCF for a reference year with an updated and more informed RCF improves accuracy.

## 4.4 Applying Model to Residence Frame and Building RCF

As described at the end of Section 2.3, the trained models are applied to the entire set of persons, $N^L$, and their residences for both the 2014 and 2015 reference years. The 2014 and 2015 RCFs have 555.7 million and 468.3 million PIK/MAFID records, respectively (see Tables 8 and 9). After scoring, the final step in production is to build the RCF and release it for use in LEHD infrastructure processes and LODES. Based on our validation results, we recommend that most downstream users apply the modal weights to calculations, whereby only a single MAFID for a person will have a weight of one and any others a weight of zero. However, users interested in representing the uncertainty in residence location should consider using the preference weights in calculations, even though overall accuracy for that weighting is a bit lower.

Table 22: Contents of RCF for Reference Year 2015

| LABEL | NAME | vartype |
|---|---|---|
| Year targeted by RCF for residence data currency | year | Num |
| Protected Indentification Key assigned to person record | pik | Char(9) |
| Master Address File ID assigned to residence record | mafid | Char(9) |
| Year history pattern of MAFID for PIK | maf_year_hist | Char(3) |
| 2010 tabulation geography STFID and block from MAF year 2016 | geocodefull | Char(15) |
| Length of geocode SSCCCTTTTTTBBBB | stfidlen | Num |
| Preference weight PDF for model 111 | pref_weight_111 | Num |
| Rank of the residential location for model 111 | pref_rank_111 | Num |
| Preference weight PDF for model 110 | pref_weight_110 | Num |
| Rank of the residential location for model 110 | pref_rank_110 | Num |
| Preference weight PDF for Rank Order model | pref_weight_carra | Num |
| Rank of residential location for Rank Order model | pref_rank_carra | Num |
| Rank of residential location for production (pref_rank_110) | pref_rank | Num |

Table 22 lists the contents of the 2015 RCF, including labels, variable names, and designations of variable type and field length. The RCF is unique by person, year, and residence, with person defined by a PIK, the reference year defined as the calendar year for which the address is valid, and the residence defined as a MAFID. The `pik`, `year`, and `mafid` variables define a unique record. For users needing the pattern of residence observations, we include `maf_year_hist`, which gives the longitudinal triad for a record in 0s and 1s centered on the reference year. For convenience, we also include the block geocode in 2010 Census tabulation geography as well as the length of the geocode in digits.

We do not include all of the weighting methods directly, but provide sufficient information to construct them. For both the "110" and "111" person-place models (see Section 4.1), we provide preference weights (`pref_weight_110` and `pref_weight_111`)

[19]The addition of residences for the third year raises the perfect and random rates and actually reduces the Achievement Index for both person-place and rank-order models compared with the "110" specification. The Acheivement Index using the "111" specification for the modal, preference, and rank-order models is 57.7, 30.8, and 35.9 for Table 20 and 60.1, 32.3, and 47.0 for Table 21.

and preference rankings (pref_rank_110 and pref_rank_111).[20] The preference weights correspond directly to the values from Equation 5, ranging from 0 to 1. The rankings prioritize records by weight, in descending order, so that the record with the highest preference weight receives a rank of 1. Subsequent records receive ascending ranks. We resolve ties arbitrarily, so that each rank appears only once for a person. Lastly, we include weights and ranks for the rank-order model prediction (pref_weight_carra and pref_rank_carra). Again, we note that these rank-order results are from our implementation and only use the CARRA rank-order. A user interested in modal weights could simply restrict the file to cases where pref_rank_110=1. For convenience of the LEHD production system, we duplicate the recommended ranking under the stable variable name pref_rank.

Table 23: Contents of RCF for Reference Year 2014

| LABEL | NAME | vartype |
|---|---|---|
| Year targeted by RCF for residence data currency | year | Num |
| Protected Indentification Key assigned to person record | pik | Char(9) |
| Master Address File ID assigned to residence record | mafid | Char(9) |
| Year history pattern of MAFID for PIK | maf_year_hist | Char(3) |
| 2010 tabulation geography STFID and block from MAF year 2016 | geocodefull | Char(15) |
| Length of geocode SSCCCTTTTTTBBBB | stfidlen | Num |
| Preference weight PDF for model 111 | pref_weight_111 | Num |
| Rank of the residential location for model 111 | pref_rank_111 | Num |
| Preference weight PDF for model 110 | pref_weight_110 | Num |
| Rank of the residential location for model 110 | pref_rank_110 | Num |
| Preference weight PDF for Rank Order model | pref_weight_carra | Num |
| Rank of residential location for Rank Order model | pref_rank_carra | Num |
| Rank of residential location for production (pref_rank_111) | pref_rank | Num |

Table 23 lists similar contents, but for the 2014 RCF. Whereas Table 22 uses data from 2014 and 2015, with pref_rank=pref_rank_110, Table 23 uses data from 2013 to 2015 (see Table 11), with pref_rank=pref_rank_111. As is shown in Section 4.3, the addition of the third year of residence data improves accuracy, so the more informed file could update the previous production cycle's release of RCF for that reference year.

# 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 basis for 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 and then overall, includes person characteristics, and is re-estimated for each year of data, will provide highly customized predictions.

The LEHD program, which already makes use of multiple imputation for demographic characteristics and workplace at multi-unit employers, 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.

There are many possibilities for enhancing the RCF. Further development of the model may including using more information on sources and residences and indicators of the quality of PIKing and geocoding. A more substantial enhancement would be to adapt the person-place model to be a job-place model, which would be ideal for use with employer-employee matched data where workers may have multiple jobs and residences and where the suitability of any residence for a job might be evaluated by the model.

# Acknowledgments

Thanks to David Brown for comments and for sharing his previous work on residential location from administrative and other sources. Thanks to Katharine Abraham for discussion and comments at the Federal Committee on Statistical Methodology (FCSM), 2015 Research Conference. 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 data staff for helping to set up the development environment.

[20] We include both models, though the second is likely to have mis-specified predictions given that no 2016 residences are available for the 2015 reference year RCF. One long term goal for RCF is to process earlier years of residence data. For earlier years, where both previous and subsequent residences will be available, the "111" model would be ideal.

# 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. In T. Dunne, J.B. Jensen, and M.J. Roberts, editors, *Producer Dynamics: New Evidence from Micro Data.*, volume 68 of *National Bureau of Economic Research Studies in Income and Wealth*, pages 149–230. Chicago: University of Chicago Press, 2009.

Fredrik Andersson, John C. Haltiwanger, Mark J. Kutzbach, Henry O. Pollakowski, and Daniel 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`.

J. David Brown. Synthesizing Numerous Pre-Existing Data Sources to Accurately Enumerate Nonresponding Decennial Housing Units. Unpublished manuscript, U.S. Census Bureau, 2013.

Melissa C. Chow, Hubert P. Janicki, Mark J. Kutzbach, Lawrence F. Warren, and Moises Yi. A Comparison of Training Modules for Administrative Records Use in Nonresponse Followup Operations: The 2010 Census and the American Community Survey. Unpublished manuscript, U.S. Census Bureau, 2016.

Matthew R. Graham, Mark J. Kutzbach, and Brian McKenzie. Design Comparison of LODES and ACS Commuting Data Products. Working Papers 14-38, Center for Economic Studies, U.S. Census Bureau, October 2014. URL `https://ideas.repec.org/p/cen/wpaper/14-38.html`.

Matthew R. Graham, Mark J. Kutzbach, and Danielle H. Sandler. Residence Candidate File (RCF), Methodology Version 0.5 - Technical Document. Unpublished manuscript, U.S. Census Bureau, 2015.

Matthew R. Graham, Mark J. Kutzbach, and Danielle 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`.

Andrew Green, Mark Kutzbach, and Lars Vilhuber. Two perspectives on commuting: A comparison of home to work flows across job-linked survey and administrative files. Unpublished manuscript, U.S. Census Bureau, 2017.

Judith K. Hellerstein, Mark J. Kutzbach, and David Neumark. Do labor market networks have an important spatial dimension? *Journal of Urban Economics*, 79:39–58, 2014. URL `http://www.sciencedirect.com/science/article/pii/S0094119013000259`.

Henry 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`.

Henry Hyatt, Erika McEntarfer, Ken Ueda, and Alexandria Zhang. Interstate migration and employer-to-employer transitions in the u.s. new evidence from administrative records data. Working Papers 16-44, Center for Economic Studies, U.S. Census Bureau, 2016. URL `ftp://ftp2.census.gov/ces/wp/2016/CES-WP-16-44.pdf`.

Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars 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`.

Sonya Rastogi, Amy O'Hara, James Noon, Ellen A. Zapata, Cindy Espinoza, Leah B Marshall, Teresa A Schellhamer, and J. David Brown. 2010 Census Match Study. 2010 Census Planning Memoranda Series No. 247, U.S. Census Bureau, 2012.

Darcy Steeg Morris, Andrew Keller, and Brian Clark. An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census. *Statistical Journal of the International Association of Official Statistics*, 32(2):177–188, 2016.

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. URL `http://econpapers.repec.org/paper/centpaper/2007-04.htm`.

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