DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 40, ARTICLE 40, PAGES 1153–1166
PUBLISHED 7 MAY 2019
http://www.demographic-research.org/Volumes/Vol40/40/
DOI: 10.4054/DemRes.2019.40.40

Research Material

IRS county-to-county migration data, 1990–2010

Mathew Hauer

James Byars

© *2019 Mathew Hauer & James Byars.*

This open-access work is published under the terms of the Creative Commons Attribution 3.0 Germany (CC BY 3.0 DE), which permits use, reproduction, and distribution in any medium, provided the original author(s) and source are given credit.
See https://creativecommons.org/licenses/by/3.0/de/legalcode

Contents

Demographic Research: Volume 40, Article 40
Research Material

IRS county-to-county migration data, 1990–2010[1]

Mathew Hauer[2]

James Byars[3]

Abstract

BACKGROUND
The county-to-county migration data of the Internal Revenue Service's (IRS) is an incredible resource for understanding migration in the United States. Produced annually since 1990 in conjunction with the US Census Bureau, the IRS migration data represents 95% to 98% of the tax-filing universe and their dependents, making the IRS migration data one of the largest sources of migration data. However, any analysis using the IRS migration data must process at least seven legacy formats of this public data across more than 2000 data files – a serious burden for migration scholars.

OBJECTIVE
To produce a single, flat data file containing complete county-to-county IRS migration flow data and to make the computer code to process the migration data freely available.

METHODS
This paper uses R to process more than 2,000 IRS migration files into a single, flat data file for use in migration research.

CONTRIBUTION
To encourage and facilitate the use of this data, we provide a single, standardized, flat data file containing county-to-county one-year migration flows for the period 1990–2010 (containing 163,883 dyadic county pairs resulting in 3.2 million county-year observations totaling over 343 million migrants) and provide the full R script to download, process, and flatten the IRS migration data.

[1] The data and code that support the creation of this data are available online at https://osf.io/wgcf3/?view_only=c5ba62fb4821421ea0621bfd0d723e61. The data resulting from this paper is available in the Supplementary Materials.

[2] Florida State University, Tallahassee, USA. Email: mehauer@fsu.edu.
[3] University of Georgia, Athens, USA.

1. Introduction

Migration flow data (i.e., the number of migrants from location *i* to location *j*) is information that is typically difficult to obtain despite its importance (Willekens et al. 2016; Rogers, Little, and Raymer 2010). Migration scholars typically focus on cross-border, international migration flow data, and recent country-to-country migration data is vital for understanding migration processes (Abel and Sander 2014; Abel 2017, 2013). However, there is growing demonstrated importance surrounding subnational migration flows (Sorichetta et al. 2016; Curtis, Fussell, and DeWaard 2015).

In the United States, subnational migration flow data is available from three primary sources, depending on the time period: the Decennial Census, the American Community Survey, and the IRS's county-to-county migration data (described in detail in the corresponding section below). The IRS migration data is a pioneering use of administrative records to estimate demographic processes and is available on an annual basis since 1990. Because of the annual availability, relatively large, long time series universe due to the administrative records, and long history of use, the IRS data is an attractive data source for conducting migration research in the United States (e.g., Curtis, Fussell, and DeWaard 2015; Molloy, Smith, and Wozniak 2011; Frey 2009). Unfortunately, this data exists in seven legacy formats, split between 2,000+ data files, making analysis with this data rather burdensome and has probably hindered the widespread adoption of this valuable resource for US migration scholarship.

To encourage and facilitate the use of this tremendous migration resource, we make two contributions: (1) We publish a single, flat, standardized data file containing all county-to-county one-year migration flows for the period 1990–2010 (containing 163,883 dyadic county pairs resulting in 3.2 million county-year observations, totaling over 343 million migrants). (2) For reproducibility, transparency, and educational purposes, we publish the open-source R code used to process the IRS data into the single, flat, standardized data file. Scholars who wish to use this data should still familiarize themselves with the strengths and weaknesses, idiosyncrasies, and design of this data (see Gross 2005; Engels and Healy 1981; Franklin and Plane 2006; Pierce 2015 for discussions on the IRS data) and with the procedures outlined in this document and in the corresponding R code.[3]

We have attempted to introduce as little postprocessing as possible to process the data into a common format. US counties are fairly stable geographic units, but some changes in county boundaries, names, and Federal Information Processing Standard Publication (FIPS) codes do occasionally occur.[4] To try and keep as close to the original data fidelity as possible, we did not recode any geographic changes and present the IRS migra-

[3] The R code used to produce this data is available in an online repository located at https://osf.io/wgcf3/?view_only=c5ba62fb4821421ea0621bfd0d723e61.

[4] FIPS is a five-digit code used to uniquely identify US counties and county equivalents.

tion data as-is. For instance, Broomfield County, Colorado (FIPS 08014), was created out of parts of Adams, Boulder, Jefferson, and Weld counties in 2001, and thus has data only after 2002. Users should be aware of any changes in county boundaries, county names, or FIPS changes that could substantially alter any analysis of this data.[5]

We organize the document as follows: First, we describe the IRS county-to-county migration data to provide an overview of the data for scholars who might be unfamiliar with the IRS migration data. Second, we provide usage notes that supply important information that may assist other researchers who want to use our data. Third, we describe our single, flat, standardized file and document important nuances in the raw IRS migration data. Finally, we describe parts of the R code used to download the IRS migration data and process it into a common format.

The IRS migration data is an incredible tool for understanding migration. By providing this data in a readily available format and the subsequent open-source computer code used to process this data, we hope to facilitate its use in descriptive, exploratory, and analytic research of migration in the United States. This data is particularly useful for understanding migration as a spatial entity and for investigating the evolution of migration systems over time.

2. IRS migration data

The IRS began using tax data to estimate migration in the 1970s and 1980s (Engels and Healy 1981; Franklin and Plane 2006) and began releasing migration data in 1990. The IRS uses individual federal tax returns, matches these individual returns between two tax years (for instance tax year 2000 and tax year 2001), and identifies migrants and nonmigrants. Beginning with tax year 1991 (migration year 1990), the IRS produces this data in conjunction with the US Census Bureau using the IRS Individual Master File, which contains every Form 1040, 1040A, and 1040EZ (Gross 2005). Migration is identified when a current year's tax form contains an address that is different from the matched preceding year's return. A nonmigrant is identified when there is no change in address between two years. For the 2002 tax year, the IRS migration data contained approximately 130 million returns (Gross 2005).

The annual series of county-to-county migration data covers 95% to 98% of the tax-filing universe (or approximately 87% of US households (Molloy, Smith, and Wozniak 2011)) and their dependents, making this data the largest migration data source for count

[5] More detailed information about county boundary, name, or FIPS changes can be found at the following locations: https://www.census.gov/geo/reference/county-changes.html, http://www.nber.org/asg/ASG_release/County_City/FIPS/FIPS_Changes.pdf, https://www.cdc.gov/nchs/data/nvss/bridged_race/County_Geography_Changes.pdf, https://www.ddorn.net/data/FIPS_County_Code_Changes.pdf.

flows between counties in the United States. The IRS derives migration information from tax filings, making those who do not file taxes most likely to be underrepresented in the migration data (Gross 2005; DeWaard, Curtis, and Fussell 2016), namely undocumented populations, the poor, the elderly, and college students (Gross 2005). However, the overwhelming majority of householders file US tax returns in the United States (Molloy, Smith, and Wozniak 2011).

The IRS reports a number of important variables in their data. They identify the origin and destination counties, the number of tax returns or filers associated with those moves (roughly analogous to the number of households and listed as the `returns` field in the raw data) who moved from county i to county j, and the number of tax exemptions associated with those moves (roughly analogous to the number of individuals and listed as the `exemptions` field in the raw data). They also report the number of nonmigrants, reported as the number of tax returns and exemptions associated with migrants from county i to county i. We treat the `exemptions` field as the total number of migrants.

Between 1990 and 2010, the IRS used the same procedures to process the county-to-county migration data. However, in 2011 the IRS introduced a new method for processing the migration data and introduced 'enhancements' to improve the overall quality of the data (Pierce 2015). The IRS introduced three major changes. First, they began basing migration on a full year of data as opposed to a partial year of data. To meet Census Bureau deadlines, the IRS processed all income tax returns filed before the end of September and did not process the returns filed between the end of September and the end of the calendar year. Beginning with migration year 2011, the IRS included the approximately 4% of returns that are filed between the end of September and December 31, allowing the IRS to produce a full calendar year's worth of migration. Second, the IRS improved the year-to-year matching, increasing the number of matched returns by 5%. Prior to 2011, the IRS used only the primary filer's taxpayer identification number (TIN), potentially excluding individuals who may be listed as a dependent in year 1 but file on their own in year 2, or in cases where a secondary filer in year 1 (such as a spouse) files as a primary filer in year 2. After 2011, the IRS broadened their matching process to include primary, secondary, and dependent TINs to improve the matching process by 5%. Third, the IRS began tabulating gross migration at the US state level by size of adjusted gross income (AGI) and the age of the primary taxpayer.

These changes to the processing of returns create a break in the historic time series. For this reason, we limit the data we process to the period 1990–2010, the last year before the new processing rules. If a scholar wishes to process any IRS migration data after 2010, the R code that we provide can be easily adapted to do so.

2.1 Comparisons to other US migration data

As stated in the preceding section, the three main sources of migration data in the United States are the Decennial Census long form, the American Community Survey, and the IRS county-to-county migration data.

Up to and including Census 2000, the long form of the Decennial Census contained the question "Where did you live five years ago?" This question provided five-year migration data once every decade. With the discontinuation of the long form in Census 2010, the Census Bureau began collecting migration information on the American Community Survey (ACS) with the question "Where did you live one year ago?" This question provides one-year migration data with each ACS release.

The Decennial long form was a robust sample, surveying approximately one in every six (16.7%) US households. The ACS is a smaller survey with a sample size of approximately 2 million US households per year. Due to the smaller sample size, the Census Bureau pools responses into five-year averages for county-to-county migration data. Thus, ACS migration data represents one-year migration data over a five-year period. The Census Bureau processes the ACS migration data and releases county-to-county migration data sets on an annual basis, reflecting the five-year average (2010–2014, 2011–2015, etc.).

The ACS migration products and the IRS migration data both have strengths and weaknesses. Table 1 compares the ACS migration products with the IRS migration data in some key areas. The ACS universe is more complete than the IRS migration universe, lacking the tax-paying universe bias present in the IRS data; however, the ACS migration data contains approximately 2% of the observations in the IRS migration data. The IRS releases the migration data annually, allowing annual comparisons while the Census Bureau suggests only nonoverlapping five-year products should be compared to each other (i.e., 2005–2009 and 2010–2014) (Brown 2009).

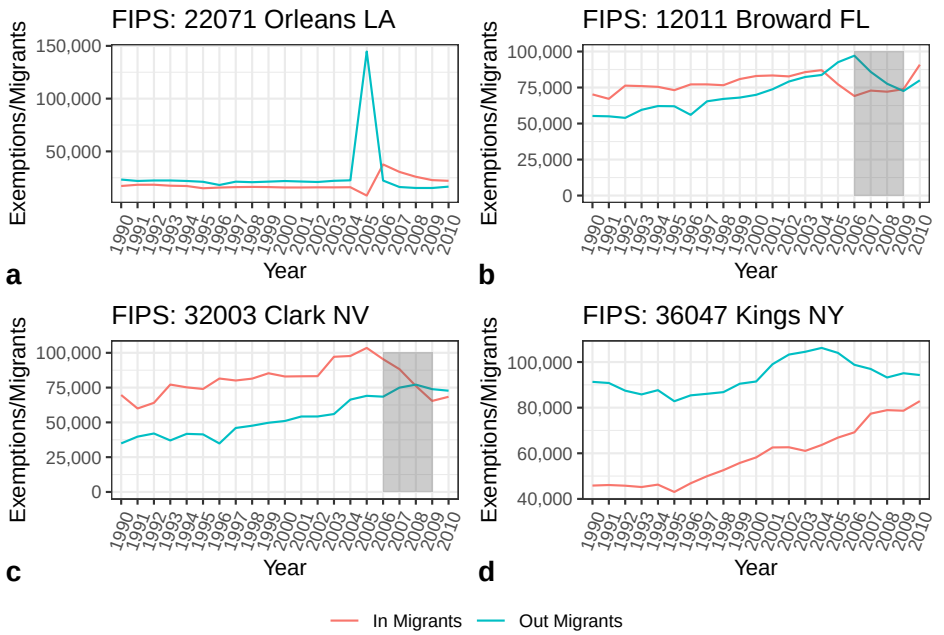
Table 1: Comparison between American Community Survey and IRS county-to-county migration data

Issue	ACS Migration Products	IRS Migration Data
Sample size	Approximately 2 million households per year	116 million+ households
Data universe	Sample is all US households	Universe is tax-filing households
Coverage period	2005–2016	1990–2016
Time period reported	Five-year average	Annual
Demographic characteristics	Each five-year product reports different sociodemographic characteristics (e.g., 2010–2014 contains relationship, household type, and tenure, 2011–2015 contains age/sex/race/Hispanic origin	No demographic characteristics

To demonstrate potential uses of the IRS migration data, Figure 1 shows detectable

changes in migration flows aggregated to gross-migration flows in four sample counties. These four sample counties are just some of the easily detectable impacts of major US events such as Hurricane Katrina in 2005 (Curtis, Fussell, and DeWaard 2015) or the Great Recession. These migration changes are largely undetectable in the ACS migration data or our ability to detect such changes is hampered by the five-year release.

Figure 1: Sample migration streams from the IRS migration data



Note: The annual release of the IRS migration data allows for detection of rapid changes in migration streams. The effect of Hurricane Katrina on New Orleans, LA, (a) is clearly visible by the large increase in out-migration in 2005. The moderate effect of the US housing bubble burst and Great Recession is detectable in Broward, FL, (b) and a much greater effect in Clark, NV, (c) as evidenced by the decline in in-migration during the 2006–2009 period. Even migration streams nearly unaffected by major US changes are also detectable, as is the case in Kings, NY, (d) which appears only marginally affected by the Great Recession or the 9/11 tragedy in 2001. These migration changes are largely undetectable in the ACS migration data or our ability to detect such changes is hampered by the five-year release. These are just a few examples of what is possible with the IRS migration data.

While the IRS migration data allows for analysis of annual changes, the IRS migration data contains no sociodemographic information. The ACS and Decennial Census migration data, on the other hand, contain county-to-county migration information crossed by sociodemographic information for some releases. If a migration scholar were interested in rapid changes in migration patterns, the IRS migration product would be

more appropriate. If a migration scholar were interested in the sociodemographic details of migration, the ACS migration products would be more appropriate.

Despite these limitations, scholars have successfully used the IRS migration data to forecast interstate migration (Isserman et al. 1985), investigate migrant destinations after Hurricane Katrina (Curtis, Fussell, and DeWaard 2015), other broader spatial patterns of migration (Henrie and Plane 2008), and to examine possible explanations (Molloy, Smith, and Wozniak 2017) to the Great Recession's migration slowdown (Frey 2009). Making this data more easily available in a more standardized format to a broader set of migration scholars can further our understanding of migration above and beyond the ACS data.

3. Usage notes

The dataset generated here provides detailed county-to-county one-year migration data based on administrative records. Users of this data should be aware that although the data has been prepared in a transparent manner with documentation of its creation and postprocessing, and with open-source computer code, little was done to postprocess the data to correct any possible inconsistencies or errors. This data should be used only with full awareness of the inherent limitations of the IRS migration data and with the knowledge of the procedures outlined in this document and in the corresponding R code. Caveat emptor – users beware.

Users should be aware of several limitations of the IRS data. Namely, that any origin-destination pair with fewer than ten tax filers is censored or suppressed by the IRS for privacy reasons. We have collected these censored flows into a unique FIPS code (FIPS 99999) by subtracting all uncensored flows from the total number of migrants. Any origin-destination pair with fewer than ten tax filers over the entire period is thus excluded from the final data file since no data would be recorded in the IRS data file due to censoring.

Users should also be mindful of possible geographic changes to county boundaries that could affect the data.

The county migration data we present comes from the `exemptions` field of the IRS migration data. The original IRS migration data contains two consistent fields across all years of data: a `returns` field and an `exemptions` field. Returns are the number of tax returns filed, while exemptions are a proxy for the members of the household. We use the number of exemptions to better mimic the number of individuals migrating rather than the number of households.

Table 2 demonstrates the general structure of our flat migration data file.

Table 2: Extract from the final migration data file

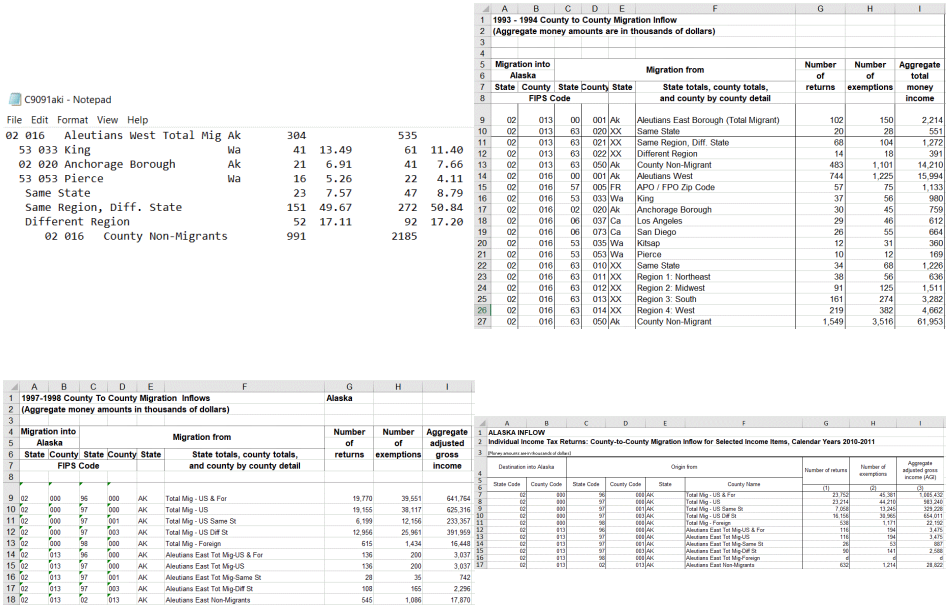
Origin	Destination	1990	1991	1992	...	2010
01001	01001	26703	27278	28677	...	40643
01001	01003	0	0	27	...	39
01001	01013	0	0	0	...	22
01001	01021	101	94	112	...	149
...	...	...	...	...	...	...
01001	99999	1324	1020	1200	...	1758

Note: Origins and Destinations are the five-digit FIPS codes, with 99999 representing all destinations with flows fewer than ten filers. The counts represent the number of `exemptions` in the IRS data. Nonmigrants are identified as having the same FIPS codes in the Origin and Destination fields.

4. Data processing

The IRS migration data for the period 1990–2010 is available in seven legacy formats. Table 3 summarizes the similarities and differences in these formats. For every year, the IRS publishes approximately 104 data files representing 52 state entities by in/out-migration for the 50 US states, DC, and a total US migration. Some years contain .csv and .dat summary files. The underlying file organization, file format, naming schema, and coding can differ between these legacy formats. Migration years 1990 and 1991 are available as fixed-width text files, while 1992–2010 are available as Microsoft Excel files. For years 1990–2003, the IRS separated in/out migration into separate folders, while 2004–2010 were published in a single folder. Each legacy format used a different file naming scheme as well, making pattern matching of file names difficult. Importantly, the IRS treated nonmigrants and total migrants differently in the seven legacy formats. For 1990 and 1991, the IRS simply had a field that read "County Non-Migrants" for nonmigrants; for 1992–1994, the IRS introduced a state code 63 but two different county codes (010 for 1992 and 1994 and 050 for 1993), creating a five-digit FIPS code of 63010 or 63050. After 1995, the IRS wisely set the origin FIPS equal to the destination FIPS for nonmigrants. Lastly, Total Migrants were treated differently too. For 1990 and 1991, the destination field simply read "Total Migration." For 1992–1994, the IRS introduced a state code 00 and county code 001 for total migrants. After 1995, the IRS used state code 96 and county code 000 for a combined five-digit FIPS code of 96000. Figure 2 shows some sample extracts of the raw IRS migration data for 1990, 1993, 1997, and 2010.

Figure 2: Sample extracts from the raw IRS migration data



C9091aki - Notepad

File Edit Format View Help

```
02 016   Aleutians West Total Mig Ak       304                535
  53 033 King                     Wa        41  13.49          61  11.40
  02 020 Anchorage Borough        Ak        21   6.91          41   7.66
  53 053 Pierce                   Wa        16   5.26          22   4.11
   Same State                              23   7.57          47   8.79
   Same Region, Diff. State               151  49.67         272  50.84
   Different Region                        52  17.11          92  17.20
      02 016   County Non-Migrants        991                2185
```

1993 - 1994 County to County Migration Inflow
(Aggregate money amounts are in thousands of dollars)

Migration into Alaska					Migration from						Number of returns	Number of exemptions	Aggregate total money income
State	County	State	County	State	State totals, county totals, and county by county detail								
FIPS Code													
02	013	00	001	Ak	Aleutians East Borough (Total Migrant)						102	150	2,214
02	013	63	020	XX	Same State						20	28	551
02	013	63	021	XX	Same Region, Diff. State						68	104	1,272
02	013	63	022	XX	Different Region						14	18	391
02	013	63	050	Ak	County Non-Migrant						483	1,101	14,210
02	016	00	001	Ak	Aleutians West						744	1,225	15,994
02	016	57	005	FR	APO / FPO Zip Code						57	75	1,133
02	016	53	033	Wa	King						37	56	980
02	016	02	020	Ak	Anchorage Borough						30	45	759
02	016	06	037	Ca	Los Angeles						29	46	612
02	016	06	073	Ca	San Diego						26	55	664
02	016	53	035	Wa	Kitsap						12	31	360
02	016	53	053	Wa	Pierce						10	12	169
02	016	63	010	XX	Same State						34	68	1,226
02	016	63	011	XX	Region 1: Northeast						38	56	636
02	016	63	012	XX	Region 2: Midwest						91	125	1,511
02	016	63	013	XX	Region 3: South						161	274	3,282
02	016	63	014	XX	Region 4: West						219	382	4,662
02	016	63	050	Ak	County Non-Migrant						1,549	3,516	61,953

1997-1998 County To County Migration Inflows — Alaska
(Aggregate money amounts in thousands of dollars)

Migration into Alaska					Migration from						Number of returns	Number of exemptions	Aggregate adjusted gross income
State	County	State	County	State	State totals, county totals, and county by county detail								
FIPS Code													
02	000	96	000	AK	Total Mig - US & For						19,770	39,551	641,764
02	000	97	000	AK	Total Mig - US						19,155	38,117	625,316
02	000	97	001	AK	Total Mig - US Same St						6,199	12,156	233,357
02	000	97	003	AK	Total Mig - US Diff St						12,956	25,961	391,959
02	000	98	000	AK	Total Mig - Foreign						615	1,434	16,448
02	013	96	000	AK	Aleutians East Tot Mig-US & For						136	200	3,037
02	013	97	000	AK	Aleutians East Tot Mig-US						136	200	3,037
02	013	97	001	AK	Aleutians East Tot Mig-Same St						28	35	742
02	013	97	003	AK	Aleutians East Tot Mig-Diff St						108	165	2,296
02	013	02	013	AK	Aleutians East Non-Migrants						545	1,086	17,870

ALASKA INFLOW
Individual Income Tax Returns: County-to-County Migration Inflow for Selected Income Items, Calendar Years 2010-2011
(Money amounts were thousands of dollars)

Destination into Alaska		Origin from				Number of returns	Number of exemptions	Aggregate adjusted gross income (AGI)
State Code	County Code	State Code	County Code	State	County Name	(1)	(2)	(3)
02	000	96	000	AK	Total Mig - US & For	23,752	45,361	1,085,420
02	000	97	000	AK	Total Mig - US	23,214	44,210	983,240
02	000	97	001	AK	Total Mig - US Same St	7,058	13,245	329,226
02	000	97	003	AK	Total Mig - US Diff St	16,156	30,965	654,011
02	000	98	000	AK	Total Mig - Foreign	538	1,151	22,192
02	013	96	000	AK	Aleutians East Tot Mig-US & For	116	194	3,475
02	013	97	000	AK	Aleutians East Tot Mig-US	116	194	3,475
02	013	97	001	AK	Aleutians East Tot Mig-Same St	29	53	887
02	013	97	003	AK	Aleutians East Tot Mig-Diff St	56	141	2,588
02	013	98	000	AK	Aleutians East Tot Mig-Foreign	6	6	66
02	013	02	013	AK	Aleutians East Non-Migrants	632	1,214	28,622

Note: Here are four sample raw data extracts for 1990, 1993, 1997, and 2010. Note that all four have different file formats, structures, and coding schemes.

Table 3 highlights additional differences that are of interest to the data we produce here. Total Migrants (i.e., FIPS 96000 for migration data after 1995) is also broken down into Total Mig – US (FIPS 97000), Total Mig – US Same State (FIPS 97001), Total Mig – US Diff St (FIPS 97003), and Total Mig – Foreign (FIPS 98000). As Table 3 shows, the IRS did not code these migration flows in this manner for all years. Additionally, the IRS treated migrants to nonspecific destinations (i.e., "Region: Midwest") differently across the entire data series. For example, in 1992 the IRS coded migrants to a different state as "FIPS 63011 for Region 1: Northeast" and provided a two-digit state identification code as "XX." By 1995, the IRS standardized these codes to either 58000 for "Other Flows – Same State" and 59001 for "Other Flows – Northeast" and provided two-digit state identification codes of "SS" for same state and "DS" for different state. The IRS treated migration to foreign locations in a similar manner.

Table 3: Select differences in the file formats, file organizations, naming, and treatment of various migration statistics

	1990–1991	1992, 1994	1993	1995–2003	2004–2006	2007–2008	2009–2010
Data format	txt	xls					
File organization	Separate in/out migration	Single folder					
Sample file naming	C9091alo.txt	C9293Alo.xls	co934alo.xls	co956alor.xls	co0405ALo.xls	co0708oAl.xls	co0910oAL.xls
Coding of nonmigrants	Destination field reads 'County Non-Migrants'	State code = 63, County code = 010	State code = 63, County code = 050	Origin FIPS = Destination FIPS			
Coding of total migrants	Destination field reads 'Total Migration'	State code = 00, County code = 001		State code = 96, County code = 000			
Coding of regional migrants	Destination field reads 'Same State,' 'Region 1,' etc.	State code = 63, County code = 010, ..., 022 State ID = XX		State code = 58 or 59, County code = 000, ..., 007, State ID = SS (Same State) or DS (Diff State)			
Coding of foreign migrants	Destination field reads 'Foreign'	State code = 57, County code = 001, ..., 007, State ID = FR		State code = 57, County code = 001, ..., 009, State ID = FR			

For simplicity and data continuity purposes, we simply create a new origin/destination (FIPS 99999) that contains all unspecified migration flows. We do this by subtracting the

number of enumerated migrants (the migration flows with greater than ten migrants) from the total number of migrants. This way, the sum of all enumerated migrants in our dataset equals the total number of migrants in the IRS dataset. And the sum of all migrants and nonmigrants for any origin in a given year should roughly approximate the county population estimate for the previous year.

The aggregation to FIPS 99999 is the only mathematical postprocessing of the IRS data.

4.1 R code

The R code used to produce this data is available in an online repository.[6] The code makes use of multicore processing to speed up computation time. There are three main sections in the code: a setup section, a data download section, and a data processing section. The final flat file, `county_migration_data.txt`, contains the # of exemptions and can be either downloaded at GitHub or produced by running the R code.

4.2 Setup

The script `000-libraries.R` simply sets up the R workspace to facilitate the data processing. The appropriate R packages are downloaded and installed if the user does not already have these packages installed. The parallel computing environment is also set up as `DetectCores() - 1` to ensure that the computer has appropriate resources for other tasks. The script requires a single reference tab-separated (tsv) file in this section, and we load it into the local environment. The `ref_state.tsv` file contains FIPS code information for US states. We simply add a FIPS state code for 'unknown' and assign it FIPS state 99.

4.3 Data download

The script `001-download_data.R` downloads and unzips the migration data from the IRS's websites and saves the formatted data in a standard file folder in the subdirectory `MigData/`. The IRS data is in two primary formats: 1990–2003 and 2004-onward. The IRS includes eight files in their zip archives that contain no data (these are in years 1998, 1999, 2000, and 2001). We delete these files after downloading and unzipping them. If they are not deleted, they cause the subsequent `for loops` to fail in the next section. These files do not contain any migration information, and their names suggest they represent aggregation of migration flows (for example 'co990usi.xls' suggests county (co)

[6] https://osf.io/wgcf3/?view_only=c5ba62fb4821421ea0621bfd0d723e61.

years 1999–2000 (990) for US (us) in-migration (i)), and we are unsure exactly why the IRS included these files or their purpose.

4.4 Data processing

The third and final section contains several `foreach` parallel processing loops to process the seven legacy formats into a common data format. These files are then row-bound using `rbindlist` and transformed into a 'short' data frame. Table 2 demonstrates the general file layout. We process the in- and out-migration files separately and keep only unique dyadic pairs in the final flat file.

References

Abel, G.J. (2013). Estimating global migration flow tables using place of birth data. *Demographic Research* 28(18): 505–546. doi:10.4054/DemRes.2013.28.18.

Abel, G.J. (2017). Estimates of global bilateral migration flows by gender between 1960 and 2015. *International Migration Review* (online). doi:10.1111/imre.12327.

Abel, G.J. and Sander, N. (2014). Quantifying global international migration flows. *Science* 343(6178): 1520–1522. doi:10.1126/science.1248676.

Brown, W.A. (2009). A compass for understanding and using American Community Survey data: What researchers need to know. Suitland: US Census Bureau.

Curtis, K.J., Fussell, E., and DeWaard, J. (2015). Recovery migration after Hurricanes Katrina and Rita: Spatial concentration and intensification in the migration system. *Demography* 52(4): 1269–1293. doi:10.1007/s13524-015-0400-7.

DeWaard, J., Curtis, K.J., and Fussell, E. (2016). Population recovery in New Orleans after Hurricane Katrina: Exploring the potential role of stage migration in migration systems. *Population and Environment* 37(4): 449–463. doi:10.1007/s11111-015-0250-7.

Engels, R.A. and Healy, M.K. (1981). Measuring interstate migration flows: An origin–destination network based on internal revenue service records. *Environment and Planning A: Economy and Space* 13(11): 1345–1360.

Franklin, R.S. and Plane, D.A. (2006). Pandora's box: The potential and peril of migration data from the American Community Survey. *International Regional Science Review* 29(3): 231–246. doi:10.1177/0160017606289895.

Frey, W. (2009). The great american migration slowdown. Washington, D.C.: Brookings Institution.

Gross, E. (2005). Internal Revenue Service area-to-area migration data: Strengths, limitations, and current trends. Washington, D.C.: Internal Revenue Service (SOI Working Paper).

Henrie, C.J. and Plane, D.A. (2008). Exodus from the California core: Using demographic effectiveness and migration impact measures to examine population redistribution within the western United States. *Population Research and Policy Review* 27(1): 43–64. doi:10.1007/s11113-007-9053-6.

Isserman, A.M., Plane, D.A., Rogerson, P.A., and Beaumont, P.M. (1985). Forecasting interstate migration with limited data: A demographic-economic approach. *Journal of the American Statistical Association* 80(390): 277–285.

Molloy, R., Smith, C.L., and Wozniak, A. (2011). Internal migration in the United States. *Journal of Economic Perspectives* 25(3): 173–196. doi:10.1257/jep.25.3.173.

Molloy, R., Smith, C.L., and Wozniak, A. (2017). Job changing and the decline in long-distance migration in the United States. *Demography* 54(2): 631–653. doi:10.1007/s13524-017-0551-9.

Pierce, K. (2015). SOI migration data: A new approach: Methodological improvements for SOI's United States population migration data, calendar years 2011–2012. Washington, D.C.: Statistics of Income, Internal Revenue Service.

Rogers, A., Little, J., and Raymer, J. (2010). *The indirect estimation of migration: Methods for dealing with irregular, inadequate, and missing data*. Dordrecht: Springer.

Sorichetta, A., Bird, T.J., Ruktanonchai, N.W., zu Erbach-Schoenberg, E., Pezzulo, C., Tejedor, N., Waldock, I.C., Sadler, J.D., Garcia, A.J., Sedda, L., and Tatem, A.J. (2016). Mapping internal connectivity through human migration in Malaria endemic countries. *Scientific Data* 3(160066): 1–16. doi:10.1038/sdata.2016.66.

Willekens, F., Massey, D., Raymer, J., and Beauchemin, C. (2016). International migration under the microscope. *Science* 352(6288): 897–899. doi:10.1126/science.aaf6545.