IRS County-to-County Migration Data *

Mathew E. Hauer [1,2*] *Florida State University*
James M. Byars [3] *University of Georgia*

*Corresponding author. mehauer@fsu.edu. p: 850-644-7103
[1] Department of Sociology, Florida State University. 605 Bellamy Building, 113 Collegiate Loop, Tallahassee, FL USA 30622.
[2] Center for Demography and Population Health, Florida State University
[3] Carl Vinson Institute of Government, University of Georgia.

BACKGROUND: The Internal Revenue Service's (IRS) county-to-county migration data are 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 represent 95 to 98 percent 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 these 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 used to process the migration data 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 migration flows for the period 1990-2010 and provide the full R script to download, process, and flatten the IRS migration data.

Introduction

Migration flow data (ie, the number of migrants from location i to location j) are typically difficult to obtain information despite their importance (Willekens et al., 2016; Rogers, Little and Raymer, 2010). Migration scholars typically focus on cross-border, national migration flow data and recent country-to-country migration data are 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 is available from three primary sources, depending on time period: the Decennial Census, the American Community Survey, and the Internal Revenue Service's (IRS) county-to-county migration data. The IRS migration data are a pioneering use of administrative records to estimate demographic processes and are available on an annual basis since 1990. Because of the annual availability, relatively long

*The data and code that supports the creation of this data are available in the Supplementary Materials and online at https://github.com/mathewhauer/IRS-migration-data.

time series, large universe due to the administrative records, and long history of use, the IRS data are an attractive data source for conducting migration research in the United States (e.g. (Curtis, Fussell and DeWaard, 2015; Molloy, Smith and Wozniak, 2011; Shumway and Otterstrom, 2001; Frey, 2009)). Unfortunately, these data exist in seven legacy formats, split between 2,000+ data files making analysis with this data rather burdensome and has likely 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 migration flows for the period 1990-2010, and (2) we publish the open-source R code used to process the IRS data into the single, flat, standardized data file for reproducibility, transparency, and educational purposes. By publishing both the migration data itself and the R code to collate, process, and flatten the IRS migration data, our hope is to save time and effort for other migration scholars and facilitate the use of this data. Scholars who wish to use these data should still familiarize themselves with the strengths and weaknesses, idiosyncrasies, and design of these 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[1].

We have attempted to introduce as little post-processing 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 FIPS codes due occasionally occur[2]. To try and keep as close to the original data fidelity as possible, we did not recode any geographic changes and present the IRS migration data as-is. For instance, Broomfield County, Colorado (FIPS 07014) was created out of parts of Adams, Boulder, Jefferson, and Weld counties in 2001 and thus has data only after 2002. Uses should be aware of any changes in county boundaries, county names, or FIPS changes that could substantially alter any analysis of this data[3].

The following document is organized 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 provide important information that may assist other researchers who 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 are an incredible tool for understanding migration. By providing these data in a readily available format and the subsequent open-source computer code used to process these data, we hope to facilitate their use in descriptive, exploratory, and analytical analyses of migration in the United States through the use of administrative data.

[1]The R code used to produce these data is available in the **Supplementary Materials** and can also be found in an online repository located at https://github.com/mathewhauer/IRS-migration-data

[2]The Federal Information Processing Standard Publication (FIPS) is a 5-digit code used to uniquely identify US counties and county equivalents.

[3]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

This is data is particularly useful for understanding migration as a spatial entity and for investigating the evolution of migration systems over time.

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 both migrants and non-migrants. Beginning with tax year 1991 (migration year 1990), the IRS produces these data in conjunction with the US Census Bureau using the IRS Individual Master File which contains every Form 1040, 1040A, and 1040EZ. Migration is identified when a current years' tax form contains an address that is different from the matched preceding years' return. A non-migrant 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 cover 95 to 98 percent of the tax filing universe (or approximately 87% of US households (Molloy, Smith and Wozniak, 2011)) and their dependents making these data the largest migration data source for count flows between counties in the United States. But since the IRS derives migration information from tax filings, those who do not file taxes are 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 both the origin and destination counties; the number of total migrants who moved from county i to county j; the number of non-migrants in origin county i; and the numbers 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) and the numbers of tax exemptions associated with those moves (roughly analogous to the number of individuals and listed as the `exemptions` field in the raw data).

Between 1990 and 2010, the IRS processed the county-to-county migration data using the same procedures. 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 that were 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 years' worth of migration. Second, the IRS improved the year-to-year matching, increasing the number of matched returns by 5 percent. 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 processes any IRS migration data after 2010, the R code that we provide can be easily adapted to do so.

Usage Notes

The dataset generated here provides detailed county-to-county migration data based on administrative records. Users of these data should be aware that although the data have been prepared in a transparent manner with documentation of their creation and post-processing, and with open-source computer code, little was done to post-process the data to correct any possible inconsistencies or errors. These 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 10 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 10 tax filers over the entire period is thus excluded from the final datafile since no data would be recorded in the IRS datafile due to censoring.

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

The county migration data we present come 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. This was done to better mimic the number of individuals migrating rather than the number of households.

Table 1 demonstrates the general structure of our flat migration data file.

Table 1: Selected file format for the final flat file. Origin and Destinations are the five-digit FIPS codes with 99999 representing all flows with fewer than 10 filers. The counts represent the number of `exemptions` in the IRS data.

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

Table 2: Select differences in the file formats, file organizations, naming, and treatment of various migration statistics.

Years	Data Format	File Organization	Sample File naming	Coding of non-migrants	Coding of Total Migrants
1990-1991	txt	Separate in/out migration	C9091alo.txt	Destination field reads 'County Non-Migrants'	Destination field reads 'Total Migration'
1992, 1994	xls	Separate in/out migration	C9293Alo.xls	State code = 63, County code = 010	State code = 00, County code = 001
1993	xls	Separate in/out migration	co934alo.xls	State code = 63, County code = 050	State code = 00, County code = 001
1995-2003	xls	Separate in/out migration	co956alor.xls	Origin FIPS = Destination FIPS	State code = 96, County = 000
2004-2006	xls	Single folder	co0405ALo.xls	Origin FIPS = Destination FIPS	State code = 96, County = 000
2007-2008	xls	Single folder	co0708oAl.xls	Origin FIPS = Destination FIPS	State code = 96, County = 000
2009-2010	xls	Single folder	co0910oAL.xls	Origin FIPS = Destination FIPS	State code = 96, County = 000

Data Processing

The IRS migration data for the period 1990-2010 are available in seven legacy formats. **Table 2** summarizes some of the similarities and differences in these formats. For every year, the IRS publishes approximately 104 data files. (52 state entities by in/out-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 excel files. For years 1990-2003, the IRS separated in/out migration into separate folders while 2004-2010 are published in a single folder. Each legacy format utilizes a different file naming scheme as well, making pattern matching of file names (called grepping) difficult. Importantly, the IRS treats non-migrants and total migrants differently in the seven legacy formats. For 1990 and 1991, the IRS simply has a field that reads "County Non-Migrants" for non-migrants; for 1992-1994, the IRS introduced a State code 63 but two difference County codes (010 for 1992 and 1994 and 050 for 1993) creating a 5-digits FIPS code of 63010 or 63050. After 1995, the IRS smartly set the origin FIPS equal to the destination FIPS for non-migrants. Lastly, Total Migrants are treated differently too. For 1990 and 1991, the destination field simply reads "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 5-digit FIPS code of 96000.

The differences described above and in **Table 2** are only some of the differences that are of interest to the data we produce here. Total Migrants, ie 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). The IRS did not code these migration flows in this manner for all years, and in some cases (such as Total Mig - Foreign) migration flows are not reported. 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 10 migrants) from the total number of migrants. This way, the sum of all enumerated migrants in our dataset will equal the total number of migrants in the IRS dataset.

The aggregation to FIPS 99999 is the only mathematical post-processing of the IRS data.

R Code

The R code used to produce these data is available in the **Supplementary Materials** and can also be found in an online repository[4]. The code makes use of multi-core 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.

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 one user core is left for other tasks that the computer might need. A single reference tab separated (tsv) file is required in this section and is loaded into the local environment. `ref_state.tsv` contains FIPS code information for US states. we simply add an additional FIPS code for 'unknown' and assign it FIPS state 99.

Data Download

The script `001-download_data.R` will download and unzip the migration data from the IRS' websites into a folder standardized format into subdirectory `MigData/`. The IRS data is in two primary formats: 1990-2003 and 2004-onward. There are eight files that the IRS includes in their zip archives that contain no data (these are in years 1998, 1999, 2000, and 2001). After being downloaded and unzipped, these files are then deleted. If they are not deleted, they will cause the subsequent `for loops` to fail in the next section. These files do not contain any migration information, their names suggest they represent aggregation of migration flows (for example 'co990usi.xls' suggests county 99-2000 US in-migration), and we are unsure exactly why these files are included or their purpose.

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 1** demonstrates the general file layout. In- and out-migration files are processed separately and only unique dyadic flows are kept in the file flat file.

[4] https://github.com/mathewhauer/IRS-migration-data

References

Abel, Guy J. 2013. "Estimating global migration flow tables using place of birth data." *Demographic Research* 28:505–546.

Abel, Guy J. 2017. "Estimates of global bilateral migration flows by gender between 1960 and 2015." *International Migration Review* .

Abel, Guy J and Nikola Sander. 2014. "Quantifying global international migration flows." *Science* 343(6178):1520–1522.

Curtis, Katherine J, Elizabeth Fussell and Jack DeWaard. 2015. "Recovery migration after Hurricanes Katrina and Rita: Spatial concentration and intensification in the migration system." *Demography* 52(4):1269–1293.

DeWaard, Jack, Katherine J Curtis and Elizabeth Fussell. 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.

Engels, Richard A and Mary K Healy. 1981. "Measuring interstate migration flows: an origin—destination network based on internal revenue service records." *Environment and Planning A* 13(11):1345–1360.

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

Frey, William. 2009. "The great American migration slowdown." *Brookings Institution, Washington, DC* .

Gross, Emily. 2005. Internal revenue service area-to-area migration data: Strengths, limitations, and current trends. In *Proceedings of the Section on Government Statistics*. p. 2005.

Molloy, Raven, Christopher L Smith and Abigail Wozniak. 2011. "Internal migration in the United States." *Journal of Economic perspectives* 25(3):173–96.

Pierce, K. 2015. "SOI migration data. A new approach: Methodological improvements for SOIC's United States population migration data, calendar years 2011–2012." *Statistics of Income, Internal Revenue Service* .

Rogers, Andrei, Jani Little and James Raymer. 2010. *The indirect estimation of migration: Methods for dealing with irregular, inadequate, and missing data*. Vol. 26 Springer Science & Business Media.

Shumway, J Matthew and Samuel M Otterstrom. 2001. "Spatial patterns of migration and income change in the Mountain West: the dominance of service-based, amenity-rich counties." *The Professional Geographer* 53(4):492–502.

Sorichetta, Alessandro, Tom J Bird, Nick W Ruktanonchai, Elisabeth zu Erbach-Schoenberg, Carla Pezzulo, Natalia Tejedor, Ian C Waldock, Jason D Sadler, Andres J Garcia, Luigi Sedda et al. 2016. "Mapping internal connectivity through human migration in malaria endemic countries." *Scientific data* 3:160066.

Willekens, Frans, Douglas Massey, James Raymer and Cris Beauchemin. 2016. "International migration under the microscope." *Science* 352(6288):897–899.