IRS Migration Database - User Summary

Janine Billadello, Geospatial Data Lab, Baruch College

May 11, 2017

Abstract

This SQLite database compiles county to county migration data produced by the IRS Statistics of Income Division (SOI) for the years 1990 - 2015. An additional state to state migration database contains migration flows for the years 1988 - 2015. When a filer submits a tax return, the address for that year is compared to the filers address in the previous year. If the two addresses differ, the filer is considered to have moved. The IRS SOI division generates inflow tables (total number of filers who moved *into* states and counties) and outflow tables (total number of filers who moved *out of* states and counties) for each year. In addition to the inflow and outflow data tables, summary tables were created with migration totals and subtotals per state and county. The annual nature of this dataset makes it valuable for the study of migration patterns within the United States over time.

Rights

Disclaimer: Every effort was made to insure that the data, which was compiled from the IRS Statistics of Income Division, was processed accurately for inclusion in the IRS Migration Database. The creator, Baruch College, and CUNY disclaim any liability for errors, inaccuracies, or omissions that may be contained therein or for any damages that may arise from the foregoing. Users should independently verify the accuracy of the data for their purposes.

IRS SOI Tax Data

The data for this project was downloaded from the IRS SOI division website:
`https://www.irs.gov/uac/soi-tax-stats-migration-data`

Most of the data has been made available in the public domain, and annual data tables continue to be published once a year. The IRS has changed their methods for collecting and tabulating the data several times over the last twenty years. The differences in raw data file formats, the sheer volume of data reported, as well as the shifting use and meaning of FIPS (Federal Information Processing Standards) codes over the years makes it challenging to use the data as a unified set. By collating the datasets into one searchable database, we are able to provide interested parties with an all-in-one resource that permits them to skip the time-consuming compilation of the datasets and focus on asking and answering questions of the data itself.

Project Goals

Given the potential of this dataset for the study of long-term migration patterns in the United States, this project collected the migration data tables into a uniform database, using Python scripts to automate the repackaging of the data into a single SQLite database. Included in the database are reference tables containing ANSI FIPS code definitions, as well as significant historical FIPS code changes that facilitate user understanding of the data over time.

SQLite IRS Migration Database

In order to browse the database, you must have a version of SQLite installed. If you are using Firefox as a web browser, SQLite is available as a free extension via the following link:
`https://addons.mozilla.org/en-us/firefox/addon/sqlite-manager/`

The SQLite database engine is available in the public domain for Windows, Mac, and Linux users via their homepage:

`https://sqlite.org/index.html`

Changes to the Data

The 2014 - 2015 data show much lower migration counts then previous years. The IRS is aware of this drop, and is investigating why it occurred.

The 2013-2014 migration data introduced a new category for the State-to-State header records. Since this category ("Total Migration - Same State") is assigned a FIPS code of 97 (the same code given to "Total Migration - US"), and appears to be a subset of the "Non-migrants" category rather then a sub-division of the true totals, it appears in the general data tables instead of in the totals tables with the other summary total headers.

A change in how the migration data was collected occurred in 2011 - 2012: the IRS began to tabulate the data based on individual income tax returns filed and received from January 1 to December 31. Previous versions (2010 - 2011 and earlier) of the migration data were based on individual income tax returns the IRS received though late September.

The table below summarizes significant changes made by the IRS in how the data was formatted, organized, or coded. These are presented by the year the change went into effect.

Year	Change to Data	Raw Data Format	IRS Summary Level Codes
1990 - 1992		Text (.txt)	
1992 - 1993	Income begins to be reported (AGI) Format change from .txt to .xls IRS summary level codes introduced	Excel (.xls)	63 - Other Flows (XX) 00 - Total Mig – US & For
1995 - 1996 (to present)	IRS summary level codes introduced		58 - Same State (SS) 59 - Different State (DS) 96 - Total Mig – US & For 97 - Tot Mig – US 98 - Tot Mig – Foreign
2004 - 2011	Format change from .xls to .dat	.dat	
2011 - 2012 (to present)	Format change from .dat to .csv	Comma-separated values (.csv)	
2013 - 2014 (to present)			97 - Total Mig – Same State

Columns in the County-level IRS Migration Database

The final database irs_migration_county.sqlite contains an Inflow and an Outflow table for each year–the following 10 columns are in each table (9 columns for 1990-91 and 1991-92, where income was not reported):

rowid	uid	st_dest_abbrv	destination	origin	st_orig_abbrv	co_orig_name	returns	exemptions	income	disclosure
1	02013_02013	AK	02013	02013	AK	Aleutians East Non-Migrants	488	955	19919	
2	02013_58000	AK	02013	58000	SS	Other Flows - Same State	30	51	952	
3	02013_59000	AK	02013	59000	DS	Other Flows - Diff State	85	149	2361	
4	02016_02016	AK	02016	02016	AK	Aleutians West Non-Migrants	1266	2633	62927	
5	02016_53033	AK	02016	53033	WA	King County	29	49	756	

Screenshot of the county Inflow table for the years 2000 - 2001

Inflow

- uid - concatenation of the destination and origin FIPS codes. This serves as a primary key for the records in each table.
- st_dest_abbrv - the two-letter state abbreviation for the state into which people are migrating.
- destination - a five-digit number representing the combined state FIPS and county FIPS codes for the county into which people are migrating (their destination county)
- origin - a five-digit number representing the combined state FIPS and county FIPS codes for the county of origin.
- st_orig_abbrv - the two-letter state abbreviation for the state in which the county of origin is located.
- co_orig_name - the name of the county of origin.
- returns - the number of the tax returns filed for a given county of origin.
- exemptions - the number of the exemptions declared by filers in a given county of origin.

- income - beginning with the 1992-93 data, this column contains the Adjusted Gross Income (AGI)–a numeric value in thousands that represents the income of filers in a given county of origin.
- disclosure - a column added to the database which contains records that have been suppressed by the IRS as a way to protect the privacy of filers from areas that had a very small number of filers in a given year; these were coded as a '-1' or as a 'd.' In the raw data, these footnotes appeared in the returns, exemptions, and income column, but were moved into the disclosure column in the database. This permits users to carry out queries on the returns, exemptions, and income data without including the -1 in their results (see *Note about disclosure and suppression*).

Note about disclosure and suppression: in cases where the number of filers falls under a certain threshold (less than 3 filers at the state-level, and less than 10 returns at the county-level), the records are not shown ("suppressed") in order to protect the confidentiality of individual filers. These records are represented by a '-1' in the data tables.

Beginning with the 2013 - 2014 data, the thresholds for inclusion in the data tabulations were raised to 10 filers for the state-level files, and 20 filers for the county-level files.

Over the years, the IRS added more summary-level categories, and accompanying codes (see the *Changes to the Data* table). These "Other Flows" categories are defined in the *Data_Changes_Definitions.docx* document. Before 1995 - 1996, these records are aggregates of suppressed records that have been combined into state (Different State or Same State) or region-level (Northeast, Midwest, South, West) categories, in order to represent migration flows at the finest possible granularity.

Starting with 1995 - 1996, the "Other Flows" categories become a breakdown of their composite category "Other Flows - Different State," and therefore appear in the totals table so as not to double count them in the general table.

For example, if less then the threshold count of filers moved from one county to another, the flow of these counties is suppressed, and the data is placed in the appropriate summary-level category. These categories illustrate that the filer moved to another county in: the same state, a different state, or a different region of the country. Further details about the suppression methods utilized by the IRS can be found in the official documentation provided by the Statistics of Income Division.

Outflow

The columns in the outflow tables are the same as the inflow tables, however the destination and origin columns are switched (since these tables show flow of migrants out of a given county and into another). The county of destination is therefore given in the co_dest_name column.

uid	st_orig_abbrv	origin	destination	st_dest_abbrv	co_dest_name	returns	exemptions	income	disclosure
53073_16001	WA	53073	16001	ID	Ada County				-1
53003_16069	WA	53003	16069	ID	Nez Perce County	203	377	3967	
53003_16057	WA	53003	16057	ID	Latah County	17	31	507	
53003_16001	WA	53003	16001	ID	Ada County	10	25	407	
53005_16001	WA	53005	16001	ID	Ada County	44	99	1442	
53005_16055	WA	53005	16055	ID	Kootenai County	27	52	917	

Screenshot of the county Outflow table for the years 1995 - 1996, showing a SQL query that returns records whose state of origin is Washington and whose destination state is Idaho. The SQL query pictured is written out here:

```
SELECT *
FROM outflow_1995_96
WHERE st_orig_abbrv = 'WA'
AND st_dest_abbrv = 'ID'
```

Reference tables included with the County-level database:

- Changes to County FIPS Codes Table ("cochanges") - A footnotes table that documents significant historical changes to the county FIPS codes over the years covered by the database. Categories of county change include the creation of new counties, the deletion or absorption of counties, and boundary changes to counties.
- FIPS Code tables ("cocodes" and "stcocodes") - A current list of the Federal Information Processing Standard (FIPS) codes is provided with the county migration database in the "cocodes" table. The "stcocode" column contains the five-digit result of the concatenation of the three-digit county FIPS code, and the two-digit state FIPS code in which a county is located. The name of the county and the abbreviation of the state it resides in are given in the "coname" and "stabbrv" columns, respectively.

The State-level IRS Migration Database

The final database "irs_migration_state.sqlite" contains an Inflow and an Outflow table for each year–the following 9 columns are in each table:

- uid - concatenation of the destination and origin FIPS codes. This serves as a primary key for the records in each table.
- st_dest_abbrv - the two-letter state abbreviation for the state into which people are migrating.
- destination - the two-digit FIPS code for the state into which people are migrating (their destination state).
- origin - the two-digit FIPS code for the state of origin.
- st_orig_abbrv - the two-letter state abbreviation for the state of origin.
- st_orig_name - the name of the state of origin.
- returns - the numeric total of the tax returns filed for a given state of origin.
- exemptions - the numeric total of the exemptions declared by filers in a given state of origin.
- income - beginning with the 1992-93 data, this column was created to hold the Adjusted Gross Income (AGI)a numeric value in thousands that represents the income of filers in a given county of origin.

A tenth column, 'disclosure' is included beginning with the 2004 - 2005 data table (see description for the County-level database).

rowid	uid	st_dest_abbrv	destination	origin	st_orig_abbrv	st_orig_name	returns	exemptions	income	disclosure
1	01_01	AL	01	01	AL	Al Non-Migrants	1501675	3401484	69502997	
2	01_13	AL	01	13	GA	Georgia	7185	15340	284234	
3	01_12	AL	01	12	FL	Florida	6242	12803	232515	
4	01_47	AL	01	47	TN	Tennessee	3176	6605	141681	
5	01_28	AL	01	28	MS	Mississippi	2727	5589	98458	

Screenshot of the state inflow table for 2004 - 2005

Reference table included with the State-level database:

A current list of the Federal Information Processing Standard (FIPS) codes is provided with the state migration database in the "stcodes" table. The "stcode" column contains the two-digit state FIPS code, "stabbrv" has the state abbreviation, and "stname" holds the full name or suppression category. There is also a "note" column that declares which years certain Total Migration or Other Flow categories were in effect.

Example Views in the IRS Migration Database

Four sample views were created to illustrate how the database can be queried to filter tables of interest, as well as generate new information from the data provided.

- flow_change_btwn_years - This view demonstrates how to compare migration flows from one year to the next, within either database. It performs an inner join on the unique ID for states or counties between two tables, and subtracts the number of tax returns filed for a given year from those of the previous year, displaying the results in a new column called "Change_Returns." If the result shown in "Change_Returns" is a negative number, it means that the Destination area had that many less filers migrate from the Origin area in the later year then they did in the year preceeding it. If the result is positive, it means that more filers moved from the Origin area to the Destination area in the later year then in the former year.
- net_change_by_year - This view uses data from the "_totals" inflow and outflow tables to derive a new column called "Net_Change." The totals tables contain the sum of all the returns, exemptions, and income rows flowing into or out of each state or county in a given tax year. Subtracting the total inflow from the total outflow yields the net change: if the resulting number is positive, it means that an area gained more people (migrants into the area) then it lost (migrants leaving the area). If the result is a negative number, it signifies that the area lost more people then it gained that year.

- inflow_single_table_pivot - This view provides an example of how the tables can be pivoted (ie. for use in a GIS software package). It filters the table to only return rows related to a single destination area, but does not include the "Non-Migrants" rows (filers who moved within the destination area itself). The example result positions the FIPS code of origin in the first column, followed by the abbreviation for the state of origin, the destination state (NY), and finally the number of returns and exemptions.

- inflow_multiarea_groupby - This view shows how multiple states or counties can be selected for a given tax year, in order to determine how many filers moved into that group of places from elsewhere. Migration between the entities in the designated group is filtered out. The resulting table is then grouped by the state or state-county FIPS code of origin, and presented in descending order by the number of tax returns filed.

  The example query for Counties focuses on the metropolitan area of New York City (five counties). Supposing someone was interested in the number of filers who moved into the NYC area from outside the area in 1995-96, the resulting table would show the county and state where those filers originated from, as well as the aggregate 'Other Flows' categories such as Same State (from other counties in New York) and Different State (from counties in states other than New York). Flows from one NYC county to another are omitted.

  The query for States returns inflow results for the New York City Tri-State area (NY, NJ, CT).

