IRS Migration Database Project Development and Workflow

By Janine Billadello May 2018

The data for this project was downloaded from <https://www.irs.gov/uac/soi-tax-stats-migration-data>. Data for migration inflow and outflow is available at the county level back to 1990-1991. The database contains county-level migration data for the years 1990-2016. State-level data goes back to 1988-1989, and the database includes state data from 1988-2016.

The raw data for this project was provided by the IRS in a variety of formats, and this document details how the files were processed. Since the format of the data tables has changed over the years, the following section will summarize the types of data tables utilized, and the scripts that were developed to handle them. Python 3.4.1 was used to collate the disparate datasets, format them in a standardized way with the appropriate FIPS code designations, and export them into a SQLite database where the data can be viewed, filtered, and queried using SQL commands. The resulting county and state IRS Migration databases contain “inflow” and “outflow” tables for each year, in addition to “totals” tables that consist of summary data for each county or state.

**Counties\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**County-level data for the years 1990-1992** were provided in a text (.txt) format for both the inflow and outflow datasets. There is a separate text file for each state in each year. For every county in a given state file, the state and county FIPS codes are provided in a ‘header line,’ with Total Number of Returns and Total Number of Exemptions into or out of that county. Indented under the header county are the migration flow counts for individual counties, these ‘sub-header’ rows contain the FIPS codes and name of the county, the state abbreviation, the Number of Returns, the percent of migration returns given as a decimal (xx.xx), the Number of Exemptions, and the percent of migration exemptions given as a decimal (xx.xx). Where applicable, undisclosed records are placed into 6-level or 3-level ‘suppression’ categories (ie. a designation of “Same State” describes a sum of suppressed migration flows into or out of a particular header county from elsewhere in the state that said county is a part of). For a record to be suppressed at the state-level, it must have had less than three returns, whereas at the county-level, a record must have had less than ten returns to be included in the suppression procedures.

Since each state file for these early years is formatted the same way, a Python script was written to loop through each text file, apply the appropriate FIPS codes for the ‘header’ county rows to the rows indented below it, and add in other FIPS identifiers for the suppressed value rows that did not have them. A decision was made to eliminate the percent returns and percent exemptions data columns, since these are not provided for years after 1991-92. The Python-formatted lists of these data tables for each year were exported to the SQLite database “irs\_migration\_county.sqlite.”

**County-level data for the years 1992-2004** were provided as Excel (.xls) spreadsheets for both the inflow and outflow datasets. There is a separate spreadsheet for each state in each year that contains all the county inflow or outflow for the counties in the state. Unlike the earlier text files, spreadsheets for these years are formatted as consistent rows, with the state and county FIPS codes for each header county provided for all the sub-counties that flow in or out of it. From 1992-93 and on, a column called “income” was added to the data tables. This Adjusted Gross Income (AGI) represents the sum total in thousands of dollars of the taxable income amounts from all applicable tax returns for a given county. Starting in these years, the IRS also began to apply the State FIPS code of 63 to the “Other Flows” categories, and county FIPS codes depending on whether the flows were part of a 6-level or 3-level suppression (see the “Data\_Changes\_Definitions.docx” file for description and list of codes).

Since the suppression categories were not assigned any FIPS codes in the early 1990-1992 files, the 63 “Other Flows” codes were retroactively applied to these years by hard-coding them into the Python script. From 1992-93 through 1994-95, the state FIPS code of 00 was assigned to the “Total Migration – US & Foreign” county header rows, in order to indicate that counties listed below them were flows into or out of that county.

As of 1995-1996, summary total categories were added for each county, as well as totals for the entire US in that year. A script was written to handle the Excel file format for these years; it read the data into lists and re-formatted it so that it was consistent with the master SQLite table structure. New SQL tables were generated for the inflow and outflow files for each state in each year, and eventually imported into the main “irs\_migration\_county.sqlite” database consisting of files from all the years provided.

**County-level data for the years 2004-2011** were provided as .dat files, while the final years 2011-12 and 2012-13 were given in .csv format. These too required a separate script to read them in and reformat them in order to have them coincide with data from the earlier tables. Formatted lists of the data from these tables were imported into the master “irs\_migration\_county.sqlite” database.

**County-level data for the years 2011-2016** were provided as .csv files. As these are the most recent years, the .csv structure of the tables is likely to continue as a standard format for IRS tax data tables in the years to come. A Python script was created based on this data structure to automate the processing of future IRS Migration datasets as they are released. Lists of the data from these tables were imported into the “irs\_migration\_county.sqlite” database.

**Notes about how the county data was processed and re-formatted:**

The database was first compiled in raw form, with all of the data imported and formatted as closely as it was to its layout in the original spreadsheets (with the exception of the deleted “percent exemptions and returns” columns for 1990-92). In order for the compilation script to run, some of the original IRS files had to have minor manual edits done to them (ie. correct the spacing of a line, or an incorrect FIPS code). For the inflow1112, inflow1213, outflow1112, and outflow1213 tables, erroneous duplicates of certain “Foreign – Other Flows” rows that showed a value of “-1” were deleted manually. A Python script identified the Unique ID numbers that appeared in the database twice, and these were filtered within the database and deleted from it using the following SQL expressions:

SELECT \* FROM [table name] WHERE Unique\_ID = [“unique id number”]

DELETE FROM inflow1112 WHERE Unique\_ID = [“unique id number”] AND Number\_Returns = "-1"

Once the all the data had been compiled into the SQLite database, another Python script was run to reformat and separate the “totals” categories from the rest of the data for each year. This script loops through each data table in succession, creates new columns that combine the state FIPS and the county FIPS codes for origin and destination into five-digit identifier codes, and adds an additional column for disclosure records, where applicable (in the final script, created to be run on future data releases, this process is condensed into one script).

The disclosure records are a way to protect the privacy of tax payers who are from areas that had a very small number of filers in a given year. When the script encounters rows that include disclosure records (these are coded “-1” or “d”), it moves them out of the exemptions and returns cell, into the disclosure cell. The script then creates the “\_totals” tables for each year, and populates these with the migration totals for each header county. This structure allows for users to run queries on the original table without risk of including the totals categories in their results (which would otherwise amount to a double-counting of records).

When deciding what codes were appropriate to add to rows that did not have FIPS codes assigned to them by the IRS (ie. the “Same State” suppression category for 1990-91), more recent data tables were consulted for codes that could be applied to the past. Over the years, more summary categories were added to the state and county tables, and the expansion of these summary data points sometimes accompanied a broad change in the FIPS codes created by the IRS (see the Yearly Summary table for significant years).

**County-level SQLite database columns:**

The final database “irs\_migration\_county.sqlite” contains an Inflow and an Outflow table for each year—these have the following 10 columns in each table (9 columns for 1990-91 and 1991-92, where income was not reported):

**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 numeric total of the tax returns filed for a given county of origin.

exemptions – the numeric total of the exemptions declared by filers in a given county 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.

disclosure – a column added to the database into which data values that are disclosed are moved, these are coded by the IRS as a -1 or d. These were identified by the script and moved into the disclosure column (rather than leaving them in the returns or exemptions column) so that simple equations could be accurately conducted on the returns or exemptions data. For counties that have disclosed values, the returns and exemptions cells become NULL, and the -1 or d appears in the disclosure column.

**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.

**Changes to County FIPS Codes Table**

The United States Census Bureau provides a list of “Substantial Changes to Counties and County Equivalent Entities,” grouped by decade: <https://www.census.gov/geo/reference/county-changes.html>

Categories of county change include the creation of new counties, the deletion or absorption of counties, and boundary changes to counties. Since many of these events necessitate that a county’s FIPS code be created, deleted, or altered—occurrences that impact the usability of the database—a footnotes table was created to document significant historical changes to the county FIPS codes over the years covered by the IRS Migration database. The resulting table is called “cochanges,” and is used as follows:

The column named “affected\_county” contains the FIPS code of the county affected by the action described. Year and date are provided for when the change took effect, if the information was provided. The “change\_code” column contains a number from 1 to 6, each of which is assigned to a generalized “change\_type” (ie. new entity, deleted entity, boundary changes…). The details of the county change are given in the “change\_text” column, and the counties that participated in the change are listed in the “participating\_county” column. For example, county 02231 (Skagway-Yakutat) was deleted in 1992, when it was absorbed into counties 02232 (Skagway-Hoonah-Angoon) and 02282 (Yakutat City). The final column, “pop\_affected,” shows the number of people in the population who were affected by a given change.

**FIPS Code Tables**

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.

State FIPS codes, the state abbreviations, and the full state names are provided in the “stcodes” table. In addition, this table contains definitions for IRS-generated state codes that were assigned to summary categories of migration data for certain years. These codes may refer to the same summary category of data, but may not be consistent across the years. In one case, the “Total Migration – US & Foreign” category, which represents an aggregate total of all migrants into or out of a state or county in a given year, was assigned the two-digit state code of “00” between the years of 1992-93 and 1994-95. Starting in 1995-96, this category began to be represented by a state FIPS code of “96,” thereby replacing the previous designation.

**States\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**State-level data for the years 1988-2004** were provided as .xls files. Separate Python scripts were written to handle the years 1988-1992 and 1992-2004, given changes in how the data tables were formatted. For each year, there is an inflow folder and an outflow folder, containing individual files for each State. As with the counties, it was decided that the “Percent of Total Migrants” columns given for the Number of Returns and Number of Exemptions be eliminated. The script also takes care of other anomalies in the tables, for instance it eliminates an unnecessary additional state abbreviation column present in certain years.

**State-level SQLite database columns:**

The final database “irs\_migration\_state.sqlite” contains Inflow and Outflow table for each year—these have the following 10 columns in each table:

**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 – 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.

disclosure – a column only added to the states database from 2004-05 and on in order to hold data values that are disclosed, which the IRS codes as a -1 or d. These were identified and moved into the disclosure column (rather than leaving them in the returns or exemptions column) so that simple equations could be accurately conducted on the returns or exemptions data. For states that have disclosed values, the returns and exemptions cells become NULL, and the -1 or d appears in the disclosure column.

**Outflow**

The column headers for the outflow tables are identical to the inflow tables, however the destination and origin columns are switched (since these tables show flow of migrants out of a given state and into another). The state of destination is therefore given in the “st\_dest\_name” column.

**FIPS Code Table**

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.