Changes to the Data and Definitions

By Janine Billadello

May 2018

This document summarizes select definitions and significant changes to data collection and processing as outlined by the broader documentation provided by the IRS SOI Tax Stats Migration Data User Guides, which can be viewed for some data years here: <https://www.irs.gov/statistics/soi-tax-stats-migration-data>

[**2013 - 2014**](https://www.irs.gov/pub/irs-soi/1314inpublicmigdoc.pdf) **Part B – Nature of Changes**

“The following changes have been made to the 2013-2014 migration data:

- The March 2016 release of the 2013-2014 migration data updates the earlier version of the data released on September 2015. The previous version did not account for revisions made to the Codebook used to geocode individual tax returns. Additionally, the previous data identified returns as migrants that should have been identified as non-migrants. Other updates include:

**A new category for the State-to-State header records (see sections E.1 and E.2)**.”

o “A new column to the Gross Migration File for the year 1 adjusted gross income (AGI)

(see section E.5)

o A new section explaining the geocoding of returns (see section C.4).

o A list of counties affected by geocoding revisions (see Appendix 1 and 2).

o Beginning with data for 2011–2012, SOI has introduced a number of enhancements to improve the data’s overall quality, as well as provide a new series of information. For more information, see “[SOI Migration Data: A New Approach](https://www.irs.gov/pub/irs-soi/soi-a-inmig-id1509.pdf).”

“In order to strengthen the disclosure protection procedures of the data, **thresholds for inclusion within the state and county tabulations have been raised to 10 (for the state files) and 20 (for the county files)**. See section D for specific details.”

“Certain tax returns have been excluded from the county-to-county files, but kept in the state-to-state files, if the return accounted for a specified percentage of a given cell. See section D for specific details.”

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[**2012 - 2013**](https://www.irs.gov/pub/irs-soi/1213inpublicmigdoc.pdf) **Part B -- Nature of Changes**

The following changes have been made to the 2012-2013 migration data (changes in how the data is collected. Compare to earlier description below):

"Beginning with the 2011-2012 file, the migration data will be based on individual income tax returns filed and received by the IRS from January 1 to December 31. Previous versions (2010-2011 and earlier) of migration data were based on individual income tax returns the IRS received through late September.

-Returns are matched on the taxpayer identification numbers of the primary, secondary, and dependent tax filers. Prior versions of the data matched returns based on the taxpayer identification number of the primary taxpayer only. See section C.2 for details.

-A Gross Migration File showing migration flows by State, levels of adjusted gross income (AGI), and age of the primary taxpayer is included. See section E.5 for specific details.

-The state-to-state and county-to-county text files (or .dat files) will no longer be provided. Instead CSV (comma separated) files will be used instead. See sections E.1.b, E.2.b, E.3.b, and E.4.b for specific details."

**Part C -- Basic Source Information**

"Migration data are based on the population of Forms 1040 that were filed and processed by the IRS during calendar years 2012 and 2013. The bulk of returns the IRS received in 2012 represent income that was earned in 2011 and the migration data correspond to returns filed for Tax Year 2011. The bulk of returns the IRS received in 2013 represent income that was earned in 2012 and the migration data correspond to returns filed for Tax Year 2012.

-For the calendar years 2012 and 2013, the bulk of returns filed with the IRS were for Tax Years 2011 (received in calendar year 2012) and Tax Year 2012 (received in calendar year 2013); however a number of individuals did file returns that represented prior tax years. For matching purposes, prior year returns are not used in the migration data.

-The address shown on the tax return is a mailing address that may not reflect the taxpayer’s actual residence. In addition, the address may not reflect the location of the taxpayer when the income was earned. A taxpayer may move after the end of the tax year but file their return on time up to nine months later from another location.

-Totals may not be comparable to other totals published elsewhere by SOI because of specific features of the migration data.[1]

-Data do not represent the full U.S. population because many individuals are not required to file an individual income tax return.

-State codes were based on the ZIP code shown on the return.

-Tax returns filed without a ZIP code and returns filed with a ZIP code that did not match the State code shown on the return were excluded.

-Tax returns where the taxpayer was claimed as a dependent on another tax return were excluded.

-Foreign tax returns as well as those filed using Army Post Office (APO) and Fleet Post Office addresses, addresses in Puerto Rico, Guam, Virgin Islands, American Samoa, Marshall Islands, Northern Marianas, and Palau have been included in the migration data

-Tax returns are assigned a State and County FIPS [2] code using a ZIP+4-to-County codebook developed by the U.S. Census Bureau.

-The age of the primary taxpayer is used to place returns in various age categories. The primary taxpayer’s age is derived by matching the Social Security numbers on the individual income tax return to information from the Social Security Administration (SSA)."

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**From countmiguser0405.txt, Part A -- Definitions and Explanations**

BASIC DATA SOURCE

"The extracts include records for individual income tax forms 1040, 1040A and 1040EZ (Beginning with the tax year 1987, the foreign category also includes forms 1040NR, 1040PR, 1040VI and 1040SS.) processed through Cycle 39 (the 39th week in the IRS's processing year) which is in late September. Returns processed after that date are not included in the data. The extracts usually contain about 95 to 98 percent of all returns filed during any particular tax year. The returns cover the tax filing units -- the filer and spouse of filer, plus all exemptions represented on the forms. The tax year 1997 file contained, for example, 105 million returns representing about 227 million persons, as defined by the exemptions reported on the forms.

Thus, there are two limitations of these data sources – file coverage and population coverage. Because the file coverage is not complete, any control counts shown in these tables will not match analogous control counts in other IRS statistical data products. Second, there are segments of the population that are not well represented by tax returns; most notably, the elderly and the poor. Thus, care should be exercised when using these data as proxies for the other population universes."

MIGRATION STATUS

"Next, migration status is determined. A non-mover is, by definition a non-migrant. However, a mover is not necessarily a migrant. That is, the tax payer may have moved, but stayed within the same state and county. The year-1 state and county geographic codes are compared to the year-2 geographic codes. If these two sets of codes are identical, then the mover is a "non-migrant;" whereas, if these codes differ, the mover is a "migrant."

NUMBER OF RETURNS

"The number of returns includes records for individual income tax forms 1040, 1040A, 1040EZ and 1040NR (the foreign category also includes forms 1040PR, 1040VI and 1040SS) processed through Cycle 39 (the 39th week in the IRS processing year) which is in late September. Returns processed after that date are not included in the data. The number of returns also exclude single returns with the filer deceased and joint returns with both the filer and spouse of filer deceased (and there are no other filer exemptions on the return); and returns that are not geographically coded. Also, the "zero exemption" returns are excluded."

NUMBER OF EXEMPTIONS

"The number of total exemptions (usually referred to as the primary/secondary less deceased method) is defined as:(a) one for the primary filer if not deceased; plus (b) one for the secondary filer if present and not deceased; plus (c) the number of children exemptions at home, away and with EIC; plus (d) the number of parents' exemptions at home or away; plus (e) the number of other exemptions. The number of exemptions is defined from the year-2 returns for all matched returns and the year-2 only returns. The number of exemptions for the year-1 only returns are by necessity, derived from the year-1 return."

**Part H -- Suppression Procedures and Part H.1 -- Suppression Procedures for the State and County Gross Migration File**

"In order to protect the confidentiality of information for individual taxpayers, data cells that are based on a small number of returns will not be shown. For state level tabulations, the cell must be based on at least 3 returns to be shown. For county level tabulations, the cell must be based on at least 10 returns to be shown. All other data cells will be suppressed. The suppression procedures are designed to maintain additivity across and within geographic levels, and comparability across data products."

-"The data cell may be suppressed by replacing the data with "d".

-"If the total number of returns is less than 10 then all data for the county will be suppressed."

Beginning with the 2013 – 2014 data, the count thresholds for inclusion within the state and county tabulations were been raised to 10 (for the state files) and 20 (for the county files).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**From 1978-92\_Codebook.pdf**

**Inflow and Outflow Years 1990 - 1991 and 1991 - 1992:**

[Refer to "1978-92\_Codebook.pdf" document for further details]

6-level suppression: 3-level suppression:

Same State Same State

Region 1: Northeast Same Region, Diff. State

Region 2: Midwest Different Region

Region 3: South

Region 4: West

Foreign

57 001 Foreign / Overseas [If the return is filed through District Office 98]

57 003 Foreign / Puerto Rico [If the return is filed through District Office 66]

57 005 Foreign / APO/FPO Zip Code [Military, if the address is an APO/FPO address

57 007 Foreign / Virgin Islands [If the return is filed through District Office 66]

Foreign appears as a summary category that may combine the above FIPS designations until the number of Returns or Exemptions is significant (at least 10), this has no assigned FIPS code.

1. Any "significant" flows are shown separately. All non-significant flows are aggregated into higher categories such as: same state, different state-region.

2. If the foreign aggregated flow is not significant, then it is added into the largest of the 4 region aggregated flows.

3. If at least one of the 4 region aggregated flows is not significant, then they are packed into 2 categories: different state-same region, and different region [3-level suppression]. "All Migration Flows" One category: all [all suppressed].

4. If at least one of these two new lines is not significant, then they are collapsed into other flows.

5. If this category is not significant, then the data are replaced with "-1"

6. If the total migration for the county is not significant, then the migration data for that county are added into another county and all data for the suppressed county are replaced with "-1."

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**For County data starting with 1992-1993, the 3 and 6-level suppressions are assigned the following codes:**

6-level suppressions: 3-level suppressions:

63 010 XX Same State 63 020 XX Same State

63 011 XX Region 1: Northeast 63 021 XX Same Region, Diff. State

63 012 XX Region 2: Midwest 63 022 XX Different Region

63 013 XX Region 3: South 63 050 XX County Non-Migrant

63 014 XX Region 4: West

63 050 XX County Non-Migrant

There is also a "Suppress All Flows" designation (63 030) which seems to mean the same thing as "All Migration Flows" in previous years--when the not-significant flows cannot be collapsed into other categories, they are aggregated into "Suppress All Flows."

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**From 'countmiguser0708.doc,' E. Summary Level Code List in the State-to-State Migration Flows Files, and F. Summary Level Code List in the County-to-County Migration Flows Files**

Total Migration 96 000 [Total Mig - US & For]

Total Migration to/from United States 97 000 [Total Mig - US]

Migration to/from different county in same state 97 001 [Total Mig - US Same St]

Migration to/from different state 97 003 [Total Mig - US Diff St]

Total Migration to/from foreign countries 98 000 [Total Mig - Foreign]

"Non-Migrants: Special codes for non-migrants are not used. The record can be identified where the origin state and county code are the same as the destination state and county code."

Migration to/from different county in same state 58 000 [Other Flows - Same State] SS

Migration to/from different state 59 000 [Other Flows - Diff State] DS

Migration to/from the Northeast region 59 001 [Other Flows - Northeast] DS

Migration to/from the Midwest region 59 003 [Other Flows – Midwest] DS

Migration to/from the South region 59 005 [Other Flows - South] DS

Migration to/from the West region 59 007 [Other Flows - West] DS

Other foreign flows 57 009 [Foreign - Other flows] FR

In the years before 1995 – 1996, the “Other Flows” categories represented unique cases where suppressed records could be summarized into Same State, Different State, or region-level groups. Regions were separate from Different State, which is where records that could not be grouped at a regional granularity were placed.

Beginning with the 1995 – 1996 data, the “Other Flows” designations became a sub-categorical breakdown of their composite category “Other Flows – Different State,” which includes them. To avoid double-counting of these records, they were moved into the totals summary tables in the database.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**From the** [**2013 - 2014**](https://www.irs.gov/pub/irs-soi/1314inpublicmigdoc.pdf) **User Guide**

Starting with the 2013 – 2014 data, the code of ‘97’ was assigned to the new category “Total Mig – Same State.”

This new category was assigned a FIPS code of 97, the same code that has historically been assigned to the “Total Migration – US” category. Although, the new “Same State” header record is amongst the Totals, the number it represents seems to be a subset of the “Non-migrants” category. In the past states files, “Non-migrants” represented a count of all filers who had not left their state. The new category breaks this apart into two rows—“Total Migration - Same State” is a count of filers who moved elsewhere in their state, while “Non-migrants” is people who filed but did not change their residence at all. For that reason, it was decided to keep this header in the general table, rather than move it to the totals table along with the other summary totals.