1-Migration_Tidy

Robert Dinterman

2016-08-17

Preparing Migration Data

The Internal Revenue Service (IRS) maintains a dataset of county-to-county level migration from 1990 to 2013. Migration data for the United States are based on year-to-year address changes reported on individual income tax returns filed with the IRS. They present migration patterns by state or by county for the entire United States and are available for inflows—the number of new residents who moved to a state or county and where they migrated from, and outflows—the number of residents leaving a state or county and where they went. The data are available for filing years 1991 through 2014 which correspond to the residence in the fiscal year of 1990 to 2013. A person is considered a migrant if they have a different filing address across two years (so the year 1990 corresponds to someone who lived in County A in 1989 and then County B in 1990). The variables within this dataset include:

- `return` - number of returns filed, which approximates the number of households that migrated.
- `exmpt` - number of personal exemptions claimed, which approximates the number of individuals.
- `agi` - total adjusted gross income, values are per $1,000s
- Aggregate migration flows at the state level, by the size of adjusted gross income (agi) and age of the primary taxpayer, starting with filing year 2011.

I have not yet parsed the data for 1990 and 1991 because the `.txt` format is not easily readable. Further, because of data concerns with a large number of negative income flows of extremely unlikely negative values for income flows, the 1992 to 1994 years are noted to be dubious:

Table 1: Inflows

year	ofips	dfips	return	exmpt	agi
1994	63010	21137	71	164	-998,560
1994	51059	51630	97	195	-995,284
1994	48027	48201	263	546	-994,258
1994	00001	21137	475	1,132	-990,239
1994	53033	53011	519	976	-979,406
1994	00001	51630	1,766	3,635	-942,328
1994	63050	48133	5,268	12,465	-881,694
1993	63050	08039	3,096	8,077	-872,193
1993	48453	48491	3,667	7,930	-870,064
1994	63050	19095	5,456	12,831	-815,647

Table 2: Outflows

year	ofips	dfips	return	exmpt	agi
1994	21001	63020	53	129	-999,190
1994	21001	00001	206	425	-996,416
1994	51059	51630	97	195	-995,284
1994	48027	48201	263	546	-994,258
1994	53033	53011	519	976	-979,406

year	ofips	dfips	return	exmpt	agi
1994	48133	63050	5,268	12,465	-881,694
1993	08039	63050	3,096	8,077	-872,193
1993	48453	48491	3,667	7,930	-870,064
1994	19095	63050	5,456	12,831	-815,647
1994	13029	63050	5,660	14,858	-803,974

where `return` *indicates aggregate tax returns between the origin county (ofips) and destination county (dfips) to proxy for number of households;* `exmpt` *refers to the number of exemptions filed which proxies population;* `agi` *is the sum of Adjusted Gross Income for all filers and is in units of $1,000s.*

Note: I will document using the inflows data although the outflows data tells a similar story.

Further Analysis of Negative Values

The negative values undoubtedly are due to some sort of clerical or data conversion error. These may still be recoverable though seeing that summary statistics are still available. The first task is to try and detect these outliers through some sort of metric. A rudimentary way of detecting is by setting an arbitrary threshold of some unrealistic amount of average negative household agi, say -$100,000 per household, and checking which values exceed this.

year	Flagged	HH_mean	POP_mean	AGI_Mean
1992	2	30	59	-7,636
1993	3	2,258	5,340	-581,550
1994	10	2,662	6,422	-918,995
1995	3	23	41	-6,644
1996	2	12	24	-1,492
1997	13	25	50	-5,320
1998	4	44	76	-7,994
1999	9	28	53	-5,932
2000	7	33	59	-4,410
2001	9	32	61	-5,582
2002	4	15	34	-2,694
2003	8	16	33	-3,956
2004	9	20	36	-6,176
2005	9	22	46	-3,354
2006	4	17	26	-3,246
2007	11	32	58	-14,916
2008	16	23	44	-5,701
2009	14	29	50	-7,762
2010	10	23	43	-7,631
2011	62	76	158	-16,837
2012	55	47	92	-12,927
2013	13	192	327	-76,209

The flagged values appear to be minimal with 277 total issues out of 2,438,004 observations. If using `agi`, then it is best to remove these flagged values. However, the `return` and `exmpt` do not appear to suffer. Using this data for population migration would be considered appropriate.

Correcting the `agi` problem is left for future work, although there are a few notes I can give from analyzing these data:

1. Appears that the large negative values are clerical errors at the county level.
2. IRS also maintains data on county income, which give total agi for all of the county filers for a given year. So these two datasets are related, although the county income data also has a problem with implausibly large negative values.
3. The agi values for a county sum up. The migration data should match the income data, if not then this should be flagged.
4. If clerical error is not at the summed level for County A (i.e. a county-to-county node is the flagged value), then one can use a constraint to back out the flagged value.
5. If there are more than one flagged value for County A in a given year, then additional constraints are necessary. For County A, this implies there are flagged values to at least County B and County C (and potentially more). The first check should be to see the number of flagged values for County B, if there are no additional flagged values then this is identified. Continue to iterate through this process for additional flags.

Correcting for these implausible numbers should be possible, however the benefit is low for this project. If income flows become a focal issue, then this will be picked up.

Suppression Issues

The IRS data are suppressed if there are fewer than 10 filers that migrate across regions. Sometimes these data are omitted completely while other times they are designated as suppressed with a -1 value, although the -1 distinction did not occur until 2004. To get an idea of how suppressed the data are, here is a quick table of suppressed values across years plus a column at the far side indicating the number of links containing the minimum number of filers (10):

year	total	Suppressed	Pct_Suppressed	Minimum	Pct_Minimum
1992	97,149	23	0.02	6,931	7.13
1993	96,746	0	0.00	6,853	7.08
1994	98,751	7	0.01	6,915	7.00
1995	106,981	6	0.01	6,817	6.37
1996	108,734	2	0.00	6,878	6.33
1997	109,319	1	0.00	6,946	6.35
1998	110,555	0	0.00	7,177	6.49
1999	111,201	0	0.00	7,109	6.39
2000	111,807	0	0.00	7,017	6.28
2001	111,581	0	0.00	7,135	6.39
2002	109,620	0	0.00	6,918	6.31
2003	109,820	0	0.00	6,890	6.27
2004	114,534	2,278	1.99	7,135	6.23
2005	118,389	2,276	1.92	7,437	6.28
2006	116,627	2,288	1.96	7,223	6.19
2007	118,691	2,256	1.90	7,495	6.31
2008	116,059	2,331	2.01	7,310	6.30
2009	110,651	2,549	2.30	6,778	6.13
2010	113,593	2,525	2.22	7,035	6.19
2011	130,092	6,425	4.94	7,898	6.07
2012	131,920	6,362	4.82	8,022	6.08
2013	85,184	8,930	10.48	0	0.00

Suppression did not occur before 2004 although it is not clear what the reasoning behind this was. For data purposes, the best practice will be to change all values which are suppressed via the -1 value to NA to delineate that we know there is a connection but do not know the specific value. This will also be helpful with the above agi problem of large, unrealistic negative values which should be coded as missing values.

We do also see that in 2013 there is a stark change in data quality and a change in the minimum number of filers. The IRS reports that:

> 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).

This change in data definitions would imply that the values before 2013 with Returns of fewer than 20 would need to be supressed in order to match the 2013 data. This is a removal of information in order to maintain stability and would depend on the particular project as to whether it may be preferrable to simply ignore 2013 and beyond.

While there may be some interesting research ideas with suppressed data, the easiest solution here is to simply drop the suppressed as there are no more than 5% of all values suppressed. We also see that the minimum threshold for reporting a migration connection (10 filers) has remained around 7,000 observations which is approximately 7% of all reported connections.

The IRS documents the changes made in the 2010–11 and 2013–14 reporting procedures on their website.

Corrections

The data are not clean or tidy, for one there is a concern of some malfunction on the IRS part because of the large negative `agi` values above. There is no correction for potentially erroneously entered/calculated data except to flag outliers. But there are other problems which can be corrected. For instance, the FIPS codes for special flows changes across time. These special flows are used to indicate a county to larger region or larger region to county relationship (i.e. Total Migrants into Wake County). The IRS is not consistent across years or within years for this distinction. Cross-checks need to be performed to ensure that all data are consistent.

Working backwards a bit, in the 2011, 2012, and 2013 documentation we have the following special codes for each county:

Summary	Abbrev	State FIPS	County FIPS
Total Mig - US & For	US	96	000
Total Mig - US	US	97	000
Total Mig - Same St	US	97	001
Total Mig - Diff St	US	97	003
Total Mig - Foreign	US	98	000
Foreign - Overseas	FR	57	001
Foreign - Puerto Rico	FR	57	003
Foreign - APO/FPO ZIPs	FR	57	005
Foreign - Virgin Islands, U.S.	FR	57	007
Foreign - Other flows	FR	57	009
Other Flows - Same State	SS	58	000
Other Flows - Diff State	DS	59	000
Other Flows - Northeast	DS	59	001
Other Flows - Midwest	DS	59	003
Other Flows - South	DS	59	005
Other Flows - West	DS	59	007

And the region codes for the associated `59` State FIPS are as follows:

Northeast Region (59-001)	Midwest Region (59-003)	South Region (59-005)	West Region (59-007)
Connecticut (09-000)	Illinois (17-000)	Alabama (01-000)	Alaska (02-000)
Maine (23-000)	Indiana (18-000)	Arkansas (05-000)	Arizona (04-000)
Massachusetts (25-000)	Iowa (19-000)	Delaware (10-000)	California (06-000)
New Hampshire(33-000)	Kansas (20-000)	D.C. (11-000)	Colorado (08-000)
New Jersey (34-000)	Michigan (26-000)	Florida (12-000)	Hawaii (15-000)
New York (36-000)	Minnesota (27-000)	Georgia (13-000)	Idaho (16-000)
Pennsylvania (42-000)	Missouri (29-000)	Kentucky (21-000)	Montana (30-000)
Rhode Island (44-000)	Nebraska (31-000)	Louisiana (22-000)	Nevada (32-000)
Vermont (50-000)	North Dakota (38-000)	Maryland (24-000)	New Mexico (35-000)
	Ohio (39-000)	Mississippi (28-000)	Oregon (41-000)
	South Dakota (46-000)	North Carolina (37-000)	Utah (49-000)
	Wisconsin (55-000)	Oklahoma (40-000)	Washington (53-000)
		South Carolina (45-000)	Wyoming (56-000)
		Tennessee (47-000)	
		Texas (48-000)	
		Virginia (51-000)	
		West Virginia (54-000)	

A further note is that a county's non-migrant population is denoted by the FIPS codes being identical (the origin and destination are the same). Continuing to work backwards, documentation before 2005 does not exist, therefore one must manually go through the files to discern how the special flows are handled. In particular, we need to figure out the non-migrants and any special codes across state borders. From what I can tell, the above is consistent for all IRS data after 1995.

A first pass is to do a quick cross-tabulation of the special codes across years. This involves subsetting the data to use `st_fips_o` values which are not equal to the FIPS codes of US States. The following chart's columns indicate the `st_fips_o` across the rows of `year` where the value in each cell is the total count of that particular `st_fips_o`:

	0	57	58	59	63	96	97	98
1992	3,145	1,815	0	0	15,606	0	0	0
1993	3,135	1,605	0	0	15,564	0	0	0

	0	57	58	59	63	96	97	98
1994	3,137	1,566	0	0	15,672	0	0	0
1995	0	1,634	3,088	9,001	0	3,180	9,460	1,089
1996	0	1,590	3,075	9,016	0	3,180	9,433	1,055
1997	0	1,562	3,077	8,954	0	3,183	9,441	1,049
1998	0	1,603	3,076	8,857	0	3,186	9,442	1,057
1999	0	1,570	3,077	8,874	0	3,182	9,440	1,055
2000	0	1,557	3,077	8,810	0	3,182	9,440	1,043
2001	0	1,603	3,084	8,977	0	3,183	9,455	1,063
2002	0	1,619	3,076	8,893	0	3,183	9,439	1,073
2003	0	1,621	3,072	9,009	0	3,178	9,426	1,073
2004	0	1,623	3,073	9,078	0	3,193	9,579	3,193
2005	0	1,637	3,076	9,253	0	3,193	9,579	3,193
2006	0	1,623	3,075	9,168	0	3,193	9,579	3,193
2007	0	1,681	3,086	9,243	0	3,194	9,582	3,194
2008	0	1,598	3,082	9,159	0	3,195	9,585	3,195
2009	0	1,215	3,078	8,911	0	3,197	9,591	3,197
2010	0	1,211	3,091	9,028	0	3,197	9,591	3,197
2011	0	3,581	3,110	15,450	0	3,193	9,576	2,514
2012	0	3,518	3,107	15,449	0	3,193	9,577	2,475
2013	0	2,931	3,117	15,396	0	3,192	9,573	2,386

Excellent. We can see a clear demarcation where, before 1995, the special codes appear to be designated with `0` and `63`. Afterwards, it appears the special codes are as documented above except with three other clear changes in data patterns:

- In 2003, total values for the `98` FIPS (Total Foreign migration) were 1073 but then increased sharply to 3193 in 2004. This change likely demonstrates that the IRS now filled in a value of 0 (or suppressed) across all counties from 2004 and beyond where before the IRS would omit a value for the county if no (or suppressed) migration with foreign areas occurred.
- In 2010, total values for the `57` FIPS (Foreign flows) were 1211 but then increased to 3581 for 2011. I suspect the same implementation occurred as above: the IRS now replaces values with 0 (or suppressed) when before they may have simply omitted any value. This can be seen with the `98` FIPS (Total - Foreign) which changed from 3197 to 2514 across the same time.
- In 2010, total values for the `59` FIPS (Other flows - across regions) were 9028 but then increased to 15450 for 2011. This represents a 0% increase in values. I suspect the same implementation occurred as above: the IRS now replaces values with 0 (or suppressed) when before they may have simply omitted any value.

The necessary corrections for handling data before 1995 involve changing the special FIPS codes involving `0` (which represent total migrants) and `63` which are more complicated:

Table 8: Pre 1995 Special Codes

	10	11	12	13	14	15	20	21	22	30	50
COUNTY NON-MIGRANT	0	0	0	0	0	0	0	0	0	0	9417
DIFFERENT REGION	0	0	0	0	0	0	0	0	3402	0	0
FOREIGN	0	0	0	0	0	565	0	0	0	0	0
REGION 1: NORTHEAST	0	5172	0	0	0	0	0	0	0	0	0
REGION 2: MIDWEST	0	0	5197	0	0	0	0	0	0	0	0

	10	11	12	13	14	15	20	21	22	30	50
REGION 3: SOUTH	0	0	0	5196	0	0	0	0	0	0	0
REGION 4: WEST	0	0	0	0	5199	0	0	0	0	0	0
SAME REGION, DIFF. STATE	0	0	0	0	0	0	0	3402	0	0	0
SAME STATE	5090	0	0	0	0	0	3402	0	0	0	0
SUPPRESS ALL FLOWS	0	0	0	0	0	0	0	0	0	800	0

The biggest change is that we need to have consistent measures of Total-Migrants and Non-Migrants. I will choose to ignore all categories which can be calculated from other values (i.e. movement within state, movement within region, etc.). All ST `63` and CTY `50` values need to be converted to the respective home region. The ST `63` and CTY `15` is loosely classified as Foreign with around 200 observations each year with an average of 11.97876 returns per observation. This indicates that the `63` "Foreign" is not likely to be movement to designated Foreign areas but rather observations that the IRS did not know what to do with and matches up well with a category called "Foreign - Other Flows."

We also have ST `57` as a Foreign category that we need to inspect. Here is a chart with the column as the `cty_fips_o` based upon being in the Foreign (`57`) category across all years:

	1	3	5	7	9
1992	571	63	1,150	31	0
1993	594	52	922	37	0
1994	809	0	757	0	0
1995	23	68	510	3	1,030
1996	14	61	511	3	1,001
1997	19	63	483	1	996
1998	20	72	506	3	1,002
1999	18	70	479	3	1,000
2000	20	73	473	3	988
2001	25	85	484	3	1,006
2002	22	94	484	4	1,015
2003	13	94	492	2	1,020
2004	31	88	483	3	1,018
2005	18	96	503	3	1,017
2006	30	117	475	4	997
2007	35	132	503	6	1,005
2008	31	129	464	6	968
2009	17	111	323	5	759
2010	22	123	319	4	743
2011	390	146	622	9	2,414
2012	408	138	586	7	2,379
2013	237	39	329	2	2,324

The foreign flows has a marked change of emphasis, from mainly classifying flows as 1 (Foreign - Overseas) before 1995 to 9 (Foreign - Other flows) for 1995 and beyond. All other maintain the same meaning and roughly the same proportions of classifications. A noted emphasis here with the `57` FIPS is that the summation of all of these equals the `98` + `000` classification (Total Migration - Foreign) that began in 1995.

As for the `58` category of "Other Flows", these are defined as:

> At the county level, counties with less than 10 returns have been aggregated into various "Other Flows" categories. The Other Flows categories are Same State, Different State, Foreign, as well as

by region (Northeast, Midwest, South, and West). See section E.6 for a list of the "Other Flows" categories and codes.

These serve the purpose of accounting up to the aggregate level. Without these miscellaneous values, then the summation across all counties for a given county may not add up to it's Total Migrants.

The following are the equivalent changes to the pre-1995 data to best sync up data all the way through:

County Total Migration Flows:	ST	CTY	Pre-1995 Change
Non-Migrant (identical FIPS)	**NA**	**NA**	**ST 63 + 50 to ST & CTY of interest**
Total Migration – US+Foreign	**96**	**000**	**ST 0 + CTY 1 to 96 + 0**
Total Migration – Foreign	**98**	**000**	**New Variable: SUM all 57 flows**
Foreign - Overseas	**57**	**001**	**Same.**
Foreign - Puerto Rico	**57**	**003**	**Same.**
Foreign - APO/FPO ZIPs	**57**	**005**	**Same.**
Foreign - Virgin Islands, U.S	**57**	**007**	**Same.**
Foreign - Other flows (FR)	**57**	**009**	**ST 63 + 15**
Total Migration – US	**97**	**000**	**97 + 0 = 96000 minus 98000**
Other Flows - Same State (SS)	58	000	ST 63 + CTY 10 & 20
Other Flows - Different State (DS)	59	000	ST 63 + CTY 21 & 22
Other Flows - Northeast (DS)	59	001	ST 63 + 11
Other Flows - Midwest (DS)	59	003	ST 63 + 12
Other Flows - South (DS)	59	005	ST 63 + 13
Other Flows - West (DS)	59	007	ST 63 + 14
Total Migration – Same State	97	001	Sum across ST XX less its own + 58000
Total Migration – Different State	97	003	97000 minus 97001

For the purposes of this project, I do not have a substantial reason to have observations based upon different states or regions. The 97001, 97003, and all ST 58 and 59 are not important to me. Future research may dictate that I need to go back and calculate these values, but for the present time there are two structures that I need to calculate:

1. County Level statistics by year: Total In-migration, Total Out-migration, and Total Non-migrants. From those values, we can construct net-migration, total population (of filers), migration rates, etc. These are the typical variables in a migration study.
2. County-to-county: two separate dataframes which only contain the pairs between counties (the IN file and OUT file). Next, diagnostic checks are to be performed to ensure that `return` and `exmpt` are closely related. If given a year, these two values match for the IN and OUT then these values are OK. If different, or if one contains a value while the other does not, then there is a problem. Easiest solution is to simply average the two then move on, but if there are large differences then these values need to be flagged and checked. This data will contain the flow from County A to County B of households, people, and income (but not likely to be used).

Starting in 1995, the IRS data also provides summaries for the flows between a state and a county with a given 000 for County Code. Because of this, one needs to take care in making their own summary statistics in a county-to-county dataset. These values will be removed because the focus is on county-to-county only.

Inflow versus Outflow Merge

The IRS data contains two sets of files: one for the inflow between regions and one for the outflow between regions. I have looked across the IRS documentation in attempt to verify that the inflow from region A to region B should also be the outflow from region B to region A. I have not found confirmation of this from documentation, so we need to attempt to figure this out through inspection of the data.

After combining the inflows and outflows files by year, origin FIPS code, and destination FIPS code, we can inspect to see how well these match. Below is a table which indicates `total` as the number of county to county pairs; `return`, `exmpt` and `agi` as the number of pairs which match for these categories; and `match` which is the percentage of pairs which correctly match out of all the observed pairs:

year	total	return	exmpt	agi	match	bad
1992	80,596	78,258	78,258	78,258	97.1%	2,338
1993	80,253	78,130	78,130	78,130	97.4%	2,123
1994	82,355	80,318	80,318	80,318	97.5%	2,037
1995	81,540	79,444	79,444	79,444	97.4%	2,096
1996	83,395	81,319	81,319	81,319	97.5%	2,076
1997	84,066	81,923	81,923	81,923	97.5%	2,143
1998	85,304	83,262	83,262	83,262	97.6%	2,042
1999	85,992	83,940	83,940	83,940	97.6%	2,052
2000	86,693	84,615	84,615	84,615	97.6%	2,078
2001	86,293	84,101	84,101	84,101	97.5%	2,192
2002	84,411	82,186	82,186	82,186	97.4%	2,225
2003	84,452	82,366	82,366	82,365	97.5%	2,086
2004	91,142	84,717	84,717	84,717	93%	6,425
2005	94,804	88,377	88,377	88,377	93.2%	6,427
2006	93,134	86,714	86,714	86,714	93.1%	6,420
2007	95,065	88,623	88,623	88,623	93.2%	6,442
2008	92,595	86,164	86,164	86,164	93.1%	6,431
2009	87,817	81,374	81,374	81,374	92.7%	6,443
2010	90,627	84,203	84,203	84,203	92.9%	6,424
2011	97,682	92,668	92,668	92,664	94.9%	5,014
2012	99,520	94,601	94,601	94,595	95.1%	4,919
2013	53,319	48,589	48,589	48,585	91.1%	4,730

There appears to be some change in the IRS procedure of reporting flows in 2004 as before the average percentage matching is greater than 97% but after this falls to below 95%. But the high percentage of match is a positive sign here. Next, we need to consider why there are values which do not match. To do so, I check to see the values where `in` are suppressed but the `out` are not suppressed and call this variable `sup_in`. I perform the opposite and term that `sup_out` and then add the two together to get `sup_total`:

year	total	sup_in	sup_out	sup_total	bad_match
1992	80,596	1,061	1,299	2,360	2.9%
1993	80,253	996	1,127	2,123	2.6%
1994	82,355	901	1,142	2,043	2.5%
1995	81,540	978	1,120	2,098	2.6%
1996	83,395	1,011	1,065	2,076	2.5%
1997	84,066	1,022	1,123	2,145	2.6%
1998	85,304	971	1,071	2,042	2.4%
1999	85,992	991	1,061	2,052	2.4%
2000	86,693	1,009	1,069	2,078	2.4%
2001	86,293	1,066	1,126	2,192	2.5%
2002	84,411	1,053	1,172	2,225	2.6%
2003	84,452	989	1,097	2,086	2.5%
2004	91,142	3,214	3,227	6,441	7.1%
2005	94,804	3,214	3,231	6,445	6.8%

year	total	sup_in	sup_out	sup_total	bad_match
2006	93,134	3,203	3,229	6,432	6.9%
2007	95,065	3,217	3,235	6,452	6.8%
2008	92,595	3,221	3,238	6,459	7%
2009	87,817	3,222	3,245	6,467	7.4%
2010	90,627	3,211	3,227	6,438	7.1%
2011	97,682	2,551	2,463	5,014	5.1%
2012	99,520	2,495	2,424	4,919	4.9%
2013	53,319	2,395	2,335	4,730	8.9%

As it turns out, the only values that do not match are due to a suppression on one side of the values but not the other. By combining the suppressed values of the inflows and outflows, we are able to get a fuller dataset for county-to-county level migration than if we only used one of these. The data are also symmetric in that the value of returns from County A to County B in the inflow and outflow are identical.

We can now move onto using these data for visualizations and modeling.