County Migration and Income - notes and links

Introduction

These maps allow users to answer questions about income and migration patterns at the county level in the United States. For example, these maps and the associated graphs let users track how many people move in and out of a particular county. Users can also analyze the incomes of immigrants and emigrants relative to the non-migrant population of the county as well as the average family size of each group.

The maps are based on the Statistics of Income (SOI) data, made available by a joint project of the Census Bureau and the Internal Revenue Service. The maps are a data visualization application that allows for quick browsing through the data and analysis of income and migration patterns at the county level.

The SOI data contain information about every county over time. To make the visualization of panel data easier, two dimensions of the data are simultaneously displayed. As the user focuses on a year-county pair, the map of the entire country shows the distribution of the variable of interest across all counties in the given year. At the same time, the graph shows how the situation in the selected county evolves over time.

Sources of data

The primary source of data is the SOI Tax Stats. The tax-based migration data had historically been a joint project undertaken by IRS and Census Bureau but it has been entirely under IRS since 2011. The data are publicly available and can be found on the IRS website (https://www.irs.gov/uac/SOI-Tax-Stats-Migration-Data). The data contain information about migration flows between US counties beginning in year 1990 and is available in separate files for each period. These files include three variables: number of tax returns, total number of exemptions on these tax returns, and total income on these tax returns. For each county, the numbers are broken down into groups that include non-migrants as well as migrants to and from other counties. Due to privacy reasons, a migration flow between two counties is reported only if the number of households migrating is 10 or greater. Otherwise, the migration between the two counties is not reported but is still included into the aggregate county migration statistic.

Migration data between years X and X+1 are based on the tax returns filed by individuals in years X and X+1. That is, the actual tax years for which incomes were reported are the X-1 and X years. Tax returns from these two years are matched by the taxpayer's identity. People who are matched and do not change the county of residence are treated as non-migrants. People who are matched and do change the county of residence are accounted for as inflow (immigration) for the destination county and outflow (emigration) for the origin county. Tax returns which are not matched (for example because the person filed tax return in only one of the two years) are not reported. The reported income and the number of exemptions are taken from the tax year X (that is returns filed in the year X+1).

Beginning with the data for the two years 2011/2012, the methodology used for matching tax returns changed. Most importantly, the period from which tax returns were collected was extended from the initial several months of the year to the entire year. This resulted in inclusion of people who file their tax returns later in the year, and who were previously excluded from the data. This group includes many families with high incomes who were previously not reported. The change in the data collection methodology can be seen by analyzing time series for counties like New York County, NY and San Francisco

County, CA. More information about the new methodology and how it affects the time series data can be found on the IRS website (https://www.irs.gov/pub/irs-soi/soi-a-inmig-id1509.pdf).

In the data for the two years 2014/2015, an apparent change in the way migration was defined was apparent. The immigration and emigration ratios for 2014-2015 were significantly lower than in past years. The median county immigration ratio dropped from 5.5 to 4.1 and the median county emigration ratio dropped from 5.7 to 4.3 between 2013 and 2014.

The SOI documentation "2014-2015, Migration Data Users Guide" is available on the IRS website at:

https://www.irs.gov/pub/irs-soi/1415inpublicmigdoc.pdf

The documentation for the 2014/2015 file notes that a change has been made that affects the migration rates. The following note has been added in section C.1, page 2.

"Due to continuing efforts to combat identity theft, the method in which the IRS processes returns may undergo changes. These processing changes may have an impact on the migration data and should be taken into account when comparing the data across years."

Communication with the IRS SOI confirms that the change in the immigration/emigration rates are due to the new procedures and are not reflective of a processing error.

However, the migration ratios for the subsequent two years 2015/2016 are consistent with the earlier ratios.

Finally, the income reported by the IRS is reported in nominal dollars. Here, the income data is rebased to real 2009 dollars using PCE price index available on the St. Louis Fed website. We use average annual values. (https://research.stlouisfed.org/fred2/series/PCEPI)

Data manipulation

Unfortunately, the data for years 1990/1991 and 1991/1992 do not have the full set of variables. Moreover, the data for years 1992/1993, 1993/1994, and 1994/1995 have slightly different format and are incompatible with the subsequent data. Therefore, the maps and graphs use only data from years 1995/1996 and thereafter. We label data obtained from the X/X+1 year pair as data in year X.

For each county-year pair (c, y) we extract nine variables: $r_{cyn}, e_{cyn}, i_{cyn}, r_{cyi}, e_{cyi}, i_{cyi}, r_{cye}, e_{cye}$, and i_{cye}, which are respectively: (a) non-migrant number of returns, (b) non-migrant number of exemptions, (c) non-migrant total income, (d) immigrant number of returns, (e) immigrant total exemptions, (f) immigrant total income, (g) emigrant number of returns, (h) emigrant total exemptions, and (i) emigrant total income for county c in year y. Denote the inverse of the price index by δ_y where y is the year, and $\delta_{2009} = 1$. Then, for each county-year pair (c, y) we calculate the following variables:

1) e_{cyn} – number of non-migrants.
2) $\delta_y i_{cyn}/e_{cyn}$ – average non-migrant income.
3) e_{cyn}/r_{cyn} – non-migrant exemptions per return.
4) $100 \times e_{cyi}/(e_{cye} + e_{cyn})$ – immigration ratio.
5) $\delta_y i_{cyi}/e_{cyi}$ – average immigrant income.
6) e_{cyi}/r_{cyi} – immigrant exemptions per return.

7) $100 \times e_{cye}/(e_{cye} + e_{cyn})$ – emigration ratio.

8) $\delta_y i_{cye}/e_{cye}$ – average emigrant income.

9) e_{cye}/r_{cye} – emigrant exemptions per return.

10) $100 \times (e_{cyi} - e_{cye})/(e_{cye} + e_{cyn})$ – net immigration ratio.

We require i) all variables to be positive with exception of (10), ii) migration ratios to be less than 1 in absolute value, iii) income levels to be below 300,000,000 dollars, and iv) exemptions per return to be between 1 and 5. Any values that do not satisfy these conditions are dropped from the dataset.

In addition, we calculate national averages in the following way:

11) $\delta_y \dfrac{\sum_{c \in USA} i_{cyn}}{\sum_{c \in USA} e_{cyn}}$ – average national non-migrant income.

12) $\delta_y \dfrac{\sum_{c \in USA} (i_{cyi} + i_{cye})}{\sum_{c \in USA} (e_{cyi} + e_{cye})}$ – average national migrant income.

13) $\dfrac{100 \times \sum_{c \in USA} (e_{cyi} + e_{cye})}{2 \sum_{c \in USA} e_{cyn} + \sum_{c \in USA} (e_{cyi} + e_{cye})}$ – average national migration ratio.

14) $\dfrac{\sum_{c \in USA} e_{cyn}}{\sum_{c \in USA} r_{cyn}}$ – average national non-migrant exemptions per return.

15) $\dfrac{\sum_{c \in USA} (e_{cyi} + e_{cye})}{\sum_{c \in USA} (r_{cyi} + r_{cye})}$ – average national migrant exemptions per return.

USA is the set of all counties. Moreover, the following medians were calculated:

16) $Q_2 \left(\left\{ \dfrac{\delta_y i_{cyn}}{e_{cyn}} : c \in USA \right\} \right)$ – median national non-migrant income.

17) $Q_2 \left(\left\{ \dfrac{\delta_y i_{cyk}}{e_{cyk}} : c \in USA, k \in \{i, e\} \right\} \right)$ – median national migrant income.

18) $Q_2 \left(\left\{ \dfrac{100 \times e_{cyi}}{e_{cye} + e_{cyn}} : c \in USA \right\} \right)$ – median national immigration rate.

19) $Q_2 \left(\left\{ \dfrac{100 \times e_{cye}}{e_{cye} + e_{cyn}} : c \in USA \right\} \right)$ – median national emigration rate.

20) $Q_2 \left(\left\{ \dfrac{e_{cyn}}{r_{cyn}} : c \in USA \right\} \right)$ – median national non-migrant exemptions per return.

21) $Q_2 \left(\left\{ \dfrac{e_{cyk}}{r_{cyk}} : c \in USA, k \in \{i, e\} \right\} \right)$ – median national migrant exemptions per return.

22) $Q_2 \left(\left\{ \dfrac{\delta_y i_{cyn}}{e_{cyn}} : c \in S_j \right\} \right)$ – median non-migrant income for state j.

23) $Q_2 \left(\left\{ \dfrac{\delta_y i_{cyi}}{e_{cyi}} : c \in S_j \right\} \right)$ – median immigrant income for state j.

24) $Q_2 \left(\left\{ \dfrac{\delta_y i_{cye}}{e_{cye}} : c \in S_j \right\} \right)$ – median emigrant income for state j.

25) $Q_2 \left(\left\{ \dfrac{100 \times e_{cyi}}{e_{cye} + e_{cyn}} : c \in S_j \right\} \right)$ – median immigration rate for state j.

26) $Q_2 \left(\left\{ \dfrac{100 \times e_{cye}}{e_{cye} + e_{cyn}} : c \in S_j \right\} \right)$ – median emigration rate for state j.

27) $Q_2 \left(\left\{ \dfrac{e_{cyn}}{r_{cyn}} : c \in S_j \right\} \right)$ – median non-migrant exemptions per return for state j.

28) $Q_2 \left(\left\{ \dfrac{e_{cyi}}{r_{cyi}} : c \in S_j \right\} \right)$ – median immigrant exemptions per return for state j.

29) $Q_2 \left(\left\{ \frac{e_{cye}}{r_{cye}} : c \in S_j \right\} \right)$ — median emigrant exemptions per return for state j.

S_j denotes the set of counties in state j. For simplicity, medians were redefined in the following way. For the ordered (from 1) set of size n, the median is the element number $\frac{n+1}{2}$ if n is odd and $\frac{n}{2} + 1$ otherwise.

Data presentation

The gradient for color-coding of the county is based on the county's quintile in the national distribution of the given variable for the given year. Counties with missing data are colored in gray.

Users can change the county of interest by clicking on it on the map. The variable of interest can be changed using the "Series" drop-down menu. Year can be changed using the "Year" drop-down menu. In most of the figures several other series are presented alongside the user's selected series. Series can be excluded by clicking on the series' name in the figure's legend.

Contact us

Please contact us at perc@tamu.edu for more information.