UNIVERSITY OF MINNESOTA

User Beware: Concerning Findings from Recent U.S. Internal Revenue Service Migration Data

Jack DeWaard†
University of Minnesota

Mathew Hauer
Florida State University

Elizabeth Fussell
Brown University

Katherine J. Curtis
University of Wisconsin-Madison

Stephan Whitaker
Federal Reserve Bank of Cleveland

Kathryn McConnell
Yale University

Kobie Price
University of Minnesota

David Egan-Robertson
University of Wisconsin-Madison

April 2020

Working Paper No. 2020-02
DOI: https://doi.org/10.18128/MPC2020-02

†Address correspondence to Jack Dewaard, University of Minnesota, 909 Social Sciences, 267 19th Ave. S., Minneapolis, MN 55455 (email: jdewaard@umn.edu). Support for this work was provided by the Minnesota Population Center at the University of Minnesota (P2C HD041023) and NSF grant SBE 1850871.

Abstract

The U.S. Internal Revenue Service (IRS) makes publicly and freely available annual place-based and place-to-place migration data at the state and county levels. Among their many uses, the IRS migration data inform estimates of net-migration as part of the U.S. Census Bureau's Population Estimates Program, which, in turn, are used for producing other annual statistics, survey design, business planning, community development programs, and federal funding allocations. In this Research Note, we document what appears to be a systemic problem with the IRS migration data since the IRS took over responsibilities for preparing these data from the U.S. Census Bureau in 2011. We conclude by speculating on possible reasons for this problem and suggesting that the post-2011 IRS migration data not be used until the IRS resolves this issue.

Keywords

Migration, Internal migration, Migration data, Internal Revenue Service, U.S. Census Bureau

Introduction and Background

The Statistics of Income (SOI) program in the U.S. Internal Revenue Service (IRS) makes publicly and freely available annual place-based and place-to-place migration data at the state and county levels (Gross 2005; Pierce 2015).[1] Relative to other publicly available sources of U.S. migration data, the IRS migration data are unique and valuable given their temporal and geographic specificity insofar as they provide annual estimates of county and county-to-county migration (DeWaard et al. 2019; Hauer and Byars 2019; Engels and Healy 1981; Isserman et al. 1982; Molloy et al. 2011). As the IRS migration data are derived from address information contained in consecutive (i.e., year-to-year) tax returns, they are also estimated to cover roughly 87 percent of all U.S. households (Molloy et al. 2011).

The principal use of the IRS migration data by the U.S. Census Bureau is to generate state and county estimates of net-migration as part of its Population Estimates Program (Toukabri 2017). Net-migration is an input into the demographic balancing equation and is used to generate intercensal population estimates, which have been shown to be very accurate (U.S. Census Bureau 2020). These population estimates are subsequently used for many purposes, including producing other annual statistics, survey design, business planning, community development programs, and federal funding allocations.

Scholarly researchers also use the IRS migration data in many applications. Early research using these data focused on describing the U.S. migration system (McHugh and Gober 1992; Plane 1987). These efforts were later expanded to examine similarities and differences in migration across U.S. regions and the rural-urban continuum (Ambinakudige and Parisi 2017; DeWaard et al. 2020; Henrie and Plane 2008; Molloy et al. 2011; Plane, Henrie, and Perry 2005; Shumway and Otterstrom 2010, 2015). The IRS migration data have also been used to study the impacts of economic shocks and incentives on migration (Coomes

[1] See https://www.irs.gov/statistics/soi-tax-stats-migration-data.

and Hoyt 2008; Vias 2010). Finally, the IRS migration data have been used to study the relationship between climate and environmental change, including extreme weather events like hurricanes and other hazards like sea level rise, and migration from and to affected states and counties (Curtis et al. 2015, 2019; DeWaard et al. 2016; Fussell et al. 2014; Hauer 2017; Shumway et al. 2014).

The IRS migration data are produced as follows (Gross 2005; Pierce 2015). First, taxpayer identification numbers (TINs) are used to match tax returns in consecutive years. Second, among matched tax returns, migrant returns are defined as those with non-matching states or counties of residence in consecutive years. Non-migrant returns are likewise defined as those with matching states or counties of residence. Third, total counts of tax returns and tax exemptions, roughly equivalent to households and individuals, respectively, and the total adjusted gross income (AGI) contained in these migrant and non-migrant returns are then tallied up at the state and county levels and disseminated.

There are four main limitations of the IRS migration data (DeWaard et al 2019, 2020; Hauer and Byars 2019). First, because these data are generated from tax returns, they exclude those who do not file a tax return. This means that groups that are less likely to file a tax return (e.g., the elderly and the poor) are underrepresented in these data. Second, these data provide limited information. The public use dataset includes only three variables: total counts of migrant and non-migrant returns (i.e., households), exemptions (i.e. individuals), and AGI at the state and county levels. Third, due to privacy concerns, the IRS county-to-county migration data include only larger flows. Before and after 2011, these data excluded small flows of less than 10 and 20 households, respectively.

The fourth limitation of the IRS migration data, which is the jumping off point for this this Research Note, is that the most recent data "are not directly comparable" with the data from prior years (Pierce 2015:2). Prior to 2011, the IRS migration data were prepared by the

U.S. Census Bureau, which, due to internal constraints and deadlines, excluded tax returns filed after the end of September each calendar year (Gross 2005). In 2011, the IRS assumed responsibility for preparing the IRS migration data and expanded the set of tax returns to include those filed by the end of December of each of calendar year (Pierce 2015). The IRS also used additional TINs—specifically, those of primary, secondary, and dependent filers—to increase match rates of tax returns in consecutive years by nearly five percent.

These sorts of comparability issues can be and frequently are managed by migration researchers when the source(s) of the discontinuities are understood. However, several strands of current research by the authors of this paper using the IRS migration data have uncovered what appears to be a systemic problem with the post-2011 data.

Approach and Results

One strand of current research by most of the authors of this paper uses the IRS migration data to study out-migration from counties impacted by the costliest hurricanes, tornadoes, and wildfires in U.S. history. In Figure 1, we display annual probabilities of household out-migration for four disaster-affected U.S. counties for year from 1990 to 2017, calculated as the number of migrant households during a given year divided by the number of households at risk of migrating at the start of the year (Bell et al. 2002).[2] Probabilities of household in-migration are also provided, with the caveat that these are not true probabilities because the risk sets, or denominators, are the populations of each of the counties shown and not the populations of the places from which households migrated. Orleans Parish, LA, and Plaquemines Parish, LA, were impacted by Hurricane Katrina in 2005 and were among the

[2] On the IRS migration data website (see Footnote 1), data files are named and organized by consecutive year (e.g., 2011-2012), which reflects the matching process used and described earlier to produce these data. Here, we refer to each data file by the first year only.

counties that experienced the greatest property losses and property losses per capita, respectively (CEMHS 2019). Jasper County, MO, was impacted by the Joplin Tornado in 2005 and experienced the greatest property losses and property losses per capita among all affected counties. Finally, the 2018 Camp Fire was largely concentrated in Paradise, CA, located in a small area in Butte County, CA.

---FIGURE 1 ABOUT HERE---

A vertical black bar is provided in each graph in Figure 1 to denote 2011, the year when the IRS took over responsibility for preparing the IRS migration data from the U.S. Census Bureau (Pierce 2015). While levels of out- and in-migration clearly differ across counties, a curiously similar trend emerges after 2011. Specifically, after 2012, out- and in-migration fall precipitously through 2014, increase dramatically through 2016, and then sharply increase or decrease thereafter. The correspondence between the levels and changes of out- and in-migration after (versus before) 2011 is also noteworthy.

We subsequently explored whether and to what extent this pattern might be indicative of systemic issue with the post-2011 IRS migration data by examining migration patterns for a random sample of four other U.S. counties: Lee County, FL, Wayne County, IL, Montgomery County, KY, and Genesse County, NY. These results are displayed in Figure 2. Here, the same patterns emerge. In each county, out- and in-migration abruptly declines after 2012 and reaches a low in 2014, increases sharply through 2016, and then declines thereafter. There is also a particularly close correspondence between out- and in-migration after 2011.

---FIGURE 2 ABOUT HERE---

Going beyond individual counties and total out- and in-migration, we calculated the Hellinger Distance (hereafter, H Distance) using the entirety of the IRS county-to-county migration data (Hauer et al. 2019; Hellinger 1909; Pardo 2005). The H Distance, *H(P,Q)*, measures the statistical distance between two discrete probability distributions, $P =$

$(p_i, \ldots, p_k)$ and $Q = (q_i, \ldots, q_k)$, and is calculated for each origin, or migrant-sending, county as follows:

$$H(P, Q) = \sqrt{1 - \sum_i^k \sqrt{p_i \times q_i}} \quad (1)$$

The probability distribution P is the set of probabilities of migrating from county i to county j in 1990, calculated from the IRS migration data. The probability distribution Q is a similar distribution for a subsequent year after 1990. Here, we calculate Q for each single year after 1990 (1991, 1992,…, 2017) relative to P (1990) to allow for a common reference point. The H Distance ranges from zero to one, with the former indicating that P and Q are identical and the latter indicating that they are the exact opposite.

As is evident in Figure 3, after 2011 and especially after 2012, both the levels of and the changes in the median H Distance are remarkably abrupt relative to earlier changes in the series. Taken together with our earlier results, this is strong evidence of what appears to be a systemic problem with the post-2011 IRS migration data.

---FIGURE 3 ABOUT HERE---

Discussion and Conclusion

The results presented in the previous section raise at least two serious questions about the post-2011 IRS migration data. First, what is the reason for the apparently systemic problem with these data? Although this problem is not acknowledged in the documentation for the post-2011 IRS migration data (Pierce 2015), two candidate explanations mentioned earlier provide viable starting points for investigation going forward: the inclusion of additional tax returns through the end of each calendar year and the use of additional TINs to increase the match rates of tax returns in consecutive years (Pierce 2015). The culprit might also involve other internal IRS processes and procedures (e.g., [changes to] the processes and procedures used to identify and exclude potentially fraudulent tax returns). Unfortunately, the IRS has

not provided documentation that acknowledges, investigates, or identifies the reason(s) for the apparent problem with the post-2011 migration data, leaving researchers to develop their own ad-hoc adjustments (Johnson et al. 2017).

The second question concerns why the post-2011 IRS migration data were publicly disseminated in the first place with the problem that we have identified in this Research Note unacknowledged and unresolved. This is important because the IRS migration data are routinely used in both scholarly and applied settings with the strong potential to affect individuals, groups and organizations, and communities in concrete ways (Toukabri 2017; U.S. Census Bureau 2020). With so much on the line, until more is known about the reasons for this apparently systemic problem with the post-2011 IRS migration data, we conclude that these data should not be used and we encourage the IRS to resolve this issue quickly and transparently.

Acknowledgements

This work is part of the projects, "Extreme Weather Disasters, Economic Losses via Migration, and Widening Spatial Inequality" and "Demographic Responses to Natural Resource Changes," funded by the National Science Foundation (Award #1850871) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development at the National Institutes of Health (Award 5R03HD095014-02), respectively. This work is also supported by center grant #P2C HD041023 awarded to the Minnesota Population Center at the University of Minnesota, center grant # P2C HD041020 awarded to the Population Studies and Training Center at Brown University, and center grant # P2C HD047873 awarded to the Center for Demography and Ecology at the University of Wisconsin-Madison by the Eunice Kennedy Shriver National Institute of Child Health and Human Development at the National Institutes of Health.

References

Ambinakudige, S. & Parisi, D. (2017). A spatiotemporal analysis of inter-county migration patterns in the United States. *Applied Spatial Analysis and Policy*, 10, 121-137.

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., & Hugo, G. (2002). Cross-national comparison of internal migration: Issues and measures. *Journal of the Royal Statistical Society A*, 165, 435-464.

CEMHS. (2019). *Spatial Hazard Events and Losses Database for the United States, Version 18.0*. Phoenix, AZ: Center for Emergency Management and Homeland Security, Arizona State University

Coomes, P. A. & Hoyt, W. H. (2008). Income taxes and the destination of movers to multistate MSAs. *Journal of Urban Economics*, 63, 920-937.

Curtis, K. J., DeWaard, J., Fussell, E., & Rosenfeld, R.A. (2019). Differential recovery migration across the rural-urban gradient: Minimal and short-term population gains for rural disaster-affected Gulf Coast counties. *Rural Sociology*, e1-e43.

Curtis, K. J., Fussell, E., & DeWaard, J. (2015). Recovery migration after Hurricanes Katrina and Rita: Spatial concentration and intensification in the migration system. *Demography*, 52, 1269-1293

DeWaard, J., Curtis, K. J., & Fussell, E. (2016). Population recovery in New Orleans after Hurricane Katrina: Exploring the potential role of stage migration in migration systems. *Population and Environment*, 37, 449-463.

DeWaard, J., Fussell, E., Curtis, K. J, & Ha, J. T. (2020). Changing spatial interconnectivity during the "Great American Migration Slowdown": A decomposition of intercounty migration rates, 1990-2010. *Population, Space and Place*, 26, e2274.

DeWaard, J., Johnson, J. E., & Whitaker, S. D. (2019). Internal migration in the United States: A comprehensive comparative assessment of the Consumer Credit Panel. *Demographic Research*, 41, 953-1006.

Engels, R.A. & Healy, M. K. (1981). Measuring interstate migration flows: An origin-destination network based on Internal Revenue Service records. *Environmental Planning A*, 13, 1345-1360.

Fussell, E., Curtis, K. J., & DeWaard, J. (2014). Recovery migration to the City of New Orleans after Hurricane Katrina: A migration systems approach. *Population and Environment*, 35, 305-322.

Gross, E. (2005). *Internal Revenue Service Area-to-Area Migration Data: Strengths, Limitations, and Current Trends*. Washington D.C: Statistics of Income Division, Internal Revenue Service.

Hauer, M. (2017). Migration induced sea-level rise could reshape the U.S. population landscape. *Nature Climate Change*, 7, 321-325.

Hauer, M. & Byars, J. (2019). IRS county-to-county migration data, 1990-2010. *Demographic Research*, 40, 1153-1166.

Hauer, M., Holloway, S.R., & Oda, T. (2019). Evacuees and migrants exhibit different migration systems after the Great East Japan Earthquake and Tsunami. Unpublished manuscript.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veründerlichen. *Journal für die reine und angewandte Mathematik*, 136, 210–271.

Henrie, C. J. & Plane, D. A. (2007). Exodus from the California core: Using demographic effectiveness and migration impact measures to examine population redistribution within the western United States. *Population Research and Policy Review*, 27, 43-64.

Isserman, A. M., Plane, D. A., & McMillen, D. B. (1982). Internal migration in the United States: An evaluation of federal data. *Review of Public Data Use*, 10, 285–311.

Johnson, K. M., Curtis, K. J., & Egan-Robertson, D. (2017). Frozen in place: Net-migration in sug-national areas of the United States in the era of the Great Recession. *Population and Development Review*, 43, 599-623.

McHugh, K. E. & Gober, P. (1992). Short-term dynamices of the U.S. interstate migration system: 1980-1988. *Growth and Change*, 23, 428-445.

Molloy, R., Smith, C. L., & Wozniak, A. (2011). Internal migration in the United States. *Journal of Economic Perspectives*, 25, 173-196.

Pardo, L. (2018). *Statistical Inference Based on Divergence Measures*. Boca Raton, FL: Chapman and Hall/CRC

Pierce, K. (2015). *SOI Migration Data, A New Approach: Methodological Improvements for SOIC's United States Population Migration Data, Calendar Years 2011-2012*. Washington D.C: Statistics of Income Division, Internal Revenue Service.

Plane, D. A. (1987). The geographic components of change in a migration system. *Geographical Analysis*, 19, 283-299.

Plane, D. A., Henrie, C. J., Perry, M. J. (2005). Migration up and down the urban hierarchy and across the life course. *Proceedings of the National Academy of Sciences of the United States of America*, 43, 15313-15318.

Shumway, J. M. & Otterstrom, S. (2010). U.S. regional income change and migration: 1995-2004. *Population, Space and Place*, 16, 483-497.

Shumway, J. M. & Otterstrom, S. (2015). Income migration and income convergence across U.S. states, 1995-2010. *Growth and Change*, 46, 593-610.

Shumway, J. M. Otterstrom, S., & Glava, S. (2014). Environmental hazards as disamenities: Selective migration and income change in the United States from 2000-2010. *Annals of the Association of American Geographers*, 104, 280-291.

Toukabri, A. (2017). *Net Migration and Population Estimates: A High Level Overview*. Washington, D.C.: U.S. Census Bureau.

U.S. Census Bureau. 2020. *Methodology for the United States Population Estimates: Vintage 2019*. Washington, D.C.: U.S. Census Bureau.

Vias, A. C. (2010). The influence of booms and busts in the U.S. economy on the interstate migration system. *Growth and Change*, 41, 115-135.

Figure 1. Annual probabilities of household migration in four extreme weather disaster-affected U.S. counties: 1990-2017

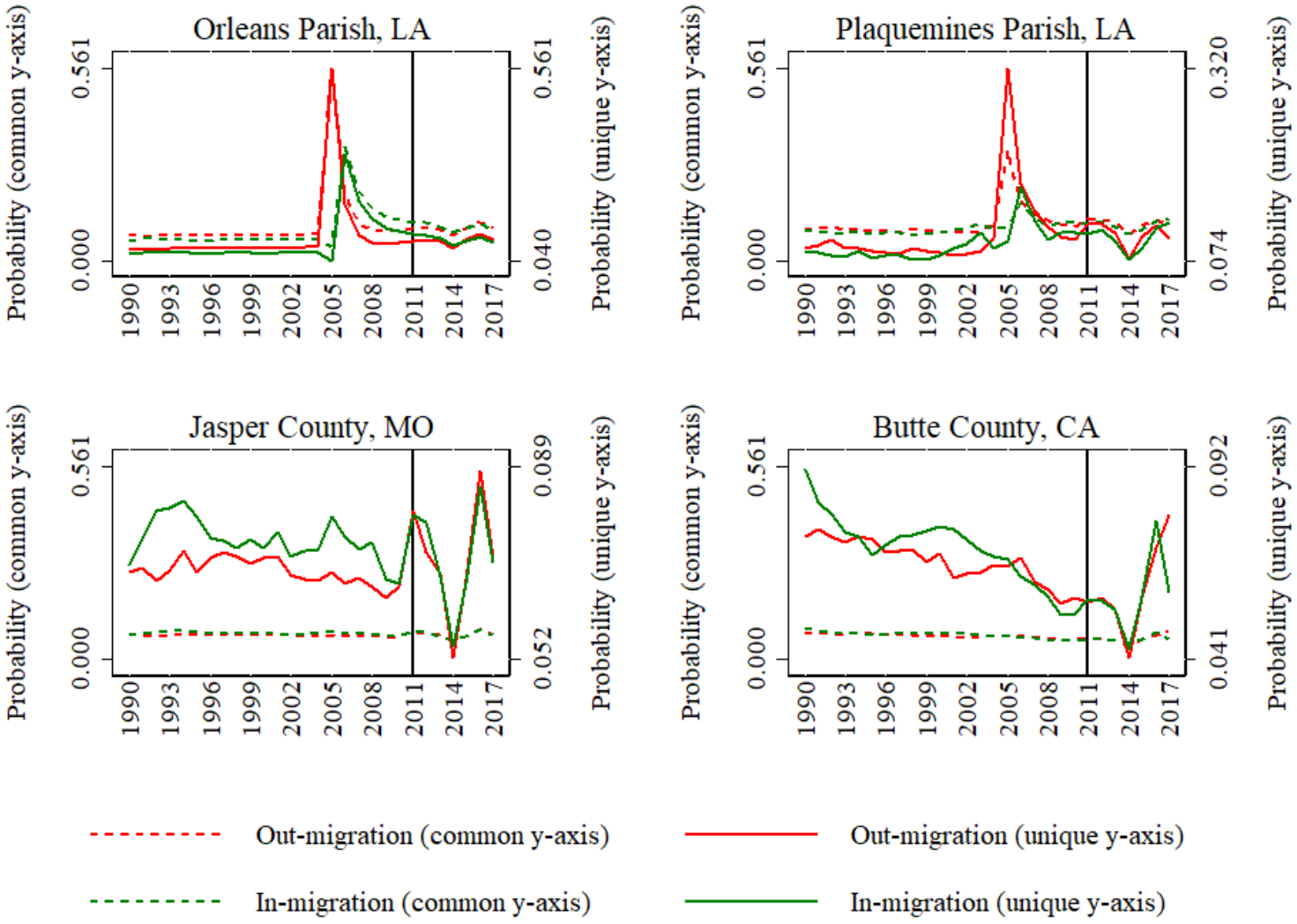

Figure 2. Annual probabilities of household migration in four randomly selected U.S. counties: 1990-2017

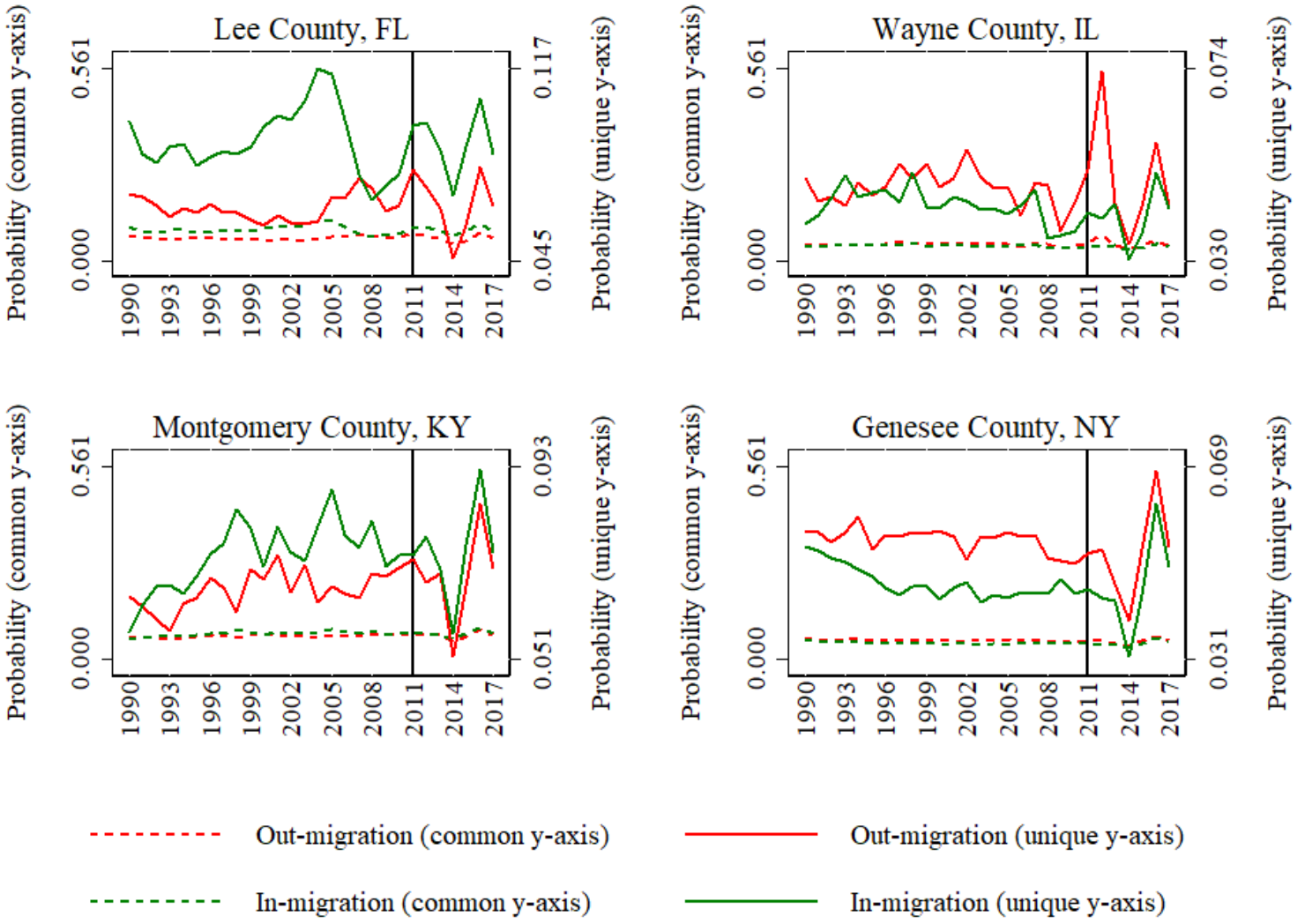

Figure 3. Hellinger (H) Distance of U.S. household county-to-county migration relative to 1990: 1991-2017