Measuring and mapping long-term changes in migration flows using population-scale family tree data

Caglar Koylu[a]* and Alice Kasakoff[b]

a Geographical and Sustainability Sciences, University of Iowa, Iowa City, USA
b Geography, University of South Carolina, Columbia, USA
* caglar-koylu@uiowa.edu

Measuring and mapping long-term changes in migration flows using population-scale family tree data

Studying migration over a long period is challenging due to lack of data, uneven data quality, and the methodological challenges that arise when mapping and analyzing migration over large geographic areas and long time spans such as changes in political boundaries. Crowd-sourced family tree data are an untapped source of volunteered geographic information generated by millions of users. These trees contain information on individuals such as birth and death places and years, and kinship ties, e.g., parent-child, spouse, and sibling relationships. Thus, family trees have the potential to support analysis of population dynamics and migration over many generations and far into the past. In this article, we introduce a methodology to measure and map long-term changes in migration flows using a population-scale family tree data set. We study internal migration in the continental United States between 1789 and 1924 using birthplaces and birthyears of children from a cleaned, geocoded and connected set of family trees from Rootsweb.com. To the best of our knowledge, the results are the first migration flow maps that show how the internal migration flows within the U.S. changed over such a long period of time (i.e., 135 years). This is one of the first studies to uncover dynamic migration patterns on a large spatial and temporal extent using family trees.

Keywords: migration, family trees, flow mapping, user-generated big data, U.S. migration history

Introduction

Studying migration over a long period of time is challenging due to many factors such as the lack of data, uneven data quality, and the methodological challenges that arise when mapping and analyzing migration over large geographic and temporal extent such as constantly changing political boundaries (Holland & Plane, 2001; Hollingsworth, 1970). The traditional source of historical migration data, the census, is available only every 10 years. Other sources, such as marriage, birth and death registries, and land records, are available at such a local level that national patterns can emerge only after a great deal of work. Crowd-sourced family tree data are an untapped source of volunteered geographic information (VGI) created and shared publicly by millions of users. Family trees include not only kinship ties (e.g., parent-child, sibling and spouse relationships) but also spatiotemporal information such as birth and death dates and locations. Thus, family trees have the potential to support analysis of population dynamics and migration in greater detail and on a large spatial and temporal scale. Several scholars have used information from such trees to generate and study the dynamics of population-scale social networks that span over many generations and far into the past (Han et al., 2017; Kaplanis et al., 2018; Kasakoff, 2019). But there have been only few attempts to study historical demography and migration using family tree data (Kandt et al., 2016; Kandt, van Dijk, & Longley, 2020; Otterstrom & Bunker, 2013; Shumway & Otterstrom, 2001).

In this article, we study internal migration in the U.S. between 1789 and 1924 using birthplaces and birthyears of children from cleaned, geocoded and connected family trees from Rootsweb.com (Koylu et al., 2020). Our main contribution is the development of a methodology to measure and map long-term changes in migration flows using population-scale family tree data. Our methodology consists of several steps. First, we extract and date birth events and location from the family tree data. We then identify family migration using the child-ladder approach which traces family migration using changes in birthplaces of consecutive children in each individual family. Next, we develop a time-series measure of the family migration rate by dividing the number of family migration events by the number of birth events. To address the uncertainty in the dating of the migration events, we employ a simple moving average window to smooth out short-term fluctuations and highlight longer-term trends in the migration rate. Next, we perform a series of algorithmic and domain-based methods to partition the study period into discrete sections to create a small number of aggregated networks that summarize a complex process of change over the 135 years. We then evaluate the changes in migration rate as well as the similarity of flows between time periods for each temporal partition to identify an optimal temporal partition. Next, we construct a time-series of migration networks using the doubly constrained gravity model and temporal normalization to account for the effect of geographic proximity, population size, and variation in the length of temporal periods. After the modularity transformation, we re-evaluate the similarity of flows between time periods and merge if a pair of consecutive time periods have similar flow patterns. Finally, we generate a series of flow maps for the optimal periodization to visualize the spatial and temporal patterns of migration. This is one of the first studies to uncover dynamic migration patterns on a larger spatial and temporal extent than the more typical micro studies of individual movement in particular localities.

The remainder of the paper is organized as follows. First, we review the background and literature on crowd-sourced family trees and historical migration studies. We then

introduce our data and describe our methodology. Finally, we present the results and conclude with a discussion of our study's findings, limitations, and future directions.

Related Work

In this section, we first introduce crowd-source family trees as a resource for studying population dynamics. We then discuss the previous work on migration and the use of family trees for demographic and migration research. Finally, we review cartographic techniques to map flows and discuss the perceptual issues and the unique challenges in migration mapping.

Crowd-sourced family tree data

Crowd-sourced family tree data are produced by millions of voluntary users who collect and organize information on their family trees using several resources such as census records, genealogy books, registries, diaries, and biographies. Users collectively link individuals across families, enter various information such as birth and death places and years, and kinship ties, which would be a daunting task for any individual or organization to address alone on this scale. However, these records often contain missing, uncertain, and duplicate information on individuals and family relationships. There is no control over accountability and accuracy of user entries. Even within the same genealogical application, trees overlap with each other, which creates multiple versions of the same families with conflicting records on events such as birth and death, and family relations such as parent-child and spouse. In our previous work, we cleaned and connected publicly available family trees from rootsweb.com to create a population-scale family tree dataset, including 250 million individuals who were born in North America and Europe between 1630 and 1930 (Koylu et al., 2020). Our methodology included data collection and cleaning, geocoding of birthplaces and deathplaces of individuals, fuzzy record linkage and a relation-based iterative search for connecting trees and deduplication of records. We evaluated the representativeness of the family tree data for population demography by comparing the individuals alive in 1880 in the U.S. from the family trees to the U.S. 1880 Census based on demographic characteristics such as gender and age, birth and death places and years of individuals and their parents.

Historical migration and family trees

Quantitative studies of migration primarily relied on historical censuses, death or burial registries, marriage licenses and contracts, because lists of actual population movements such as settlement certificates for emigrant are rather limited in scope (Hollingsworth, 1970). Previous studies mapped migration using a single source such as census records, registries, and ethnographic studies (Bertin, 2010; Dorling, 1998; Holland & Plane, 2001; Szego, 1987; Tobler, 1987), while family trees include multiple sources of data. The U.S. census, which asked about the state of birth beginning in 1850, provide information for the entire population but only in static snapshots of the populations at 10-year intervals. The recent digitization and public availability of more and more historical sources through genealogy websites have made it possible for many people to compile and share their family trees. But only a few researchers have used large data sets of user-contributed family trees to study social processes (Han et al., 2017; Kaplanis et al., 2018; Otterstrom & Bunker, 2013; Pooley & Turnbull, 1997; Price et

al., 2021), and much of the research has been genetic. Nelson (2020) introduced a methodology to estimate patrilineal kin propinquity using the sequential ordering of households in the census and used this method to identify the historical change in kin propinquity between 1800 and 1940 using historical census data. Nelson's (2020) findings revealed that the decline in kin propinquity was influenced by a number of factors such as urbanization, the decline of agriculture and kin availability, growing distance between potential kin links, and change in preferences for living near kin. Effect of nearby kin has strong influence upon migration choices, and the change in kin propinquity over time is one of the fundamental processes that determine temporal patterns of flows. While some historical studies have analyzed family ties, their focus was limited to particular kinds of relatives (Nelson, 2020) or specific areas (Otterstrom & Bunker, 2013). None has studied a large country such as the U.S. throughout its history as migrants from different origins arrived in successive waves and put down roots.

Flow mapping in space and time

Thanks to the increased use of GPS-enabled devices and sensors, movement trajectory data have increasingly become available in diverse application domains such as human mobility, and movement ecology. Individual movement data are often aggregated and anonymized into origin-destination (OD) flows given a time period. OD flow maps are commonly used to visualize movement (flows) and facilitate the understanding of patterns of flows (Tobler, 1987). To simplify flow map displays, a number of methods such as filtering (Beecham & Wood, 2014), edge bundling (Holten & van Wijk, 2009), location clustering (Adrienko & Adrienko, 2011), flow-based regionalization (Guo, 2009), and flow data smoothing and clustering methods (Zhu et al., Guo & Zhu, 2014; 2019) have been introduced. Geographically embedded matrix visualization such as the OD map (Wood, Dykes, & Slingsby, 2010) and ring maps (Zhao, Forer, & Harvey, 2008) are alternative methods to flow maps, which avoid the visual cluttering.

Existing methods of space-time cartography such as time-series maps and glyphs, change maps (Slocum et al., 2009), map animation (Fish, Goldsberry, & Battersby, 2011; Lobben, 2003; Tversky, Morrison, & Betrancourt, 2002) and space-time cubes (Hägerstrand, 1976; Kraak, 2003) could be used to identify and visualize temporal changes in spatial flow patterns. However, only a few studies have attempted to do this (Andrienko et al., 2017; Ilya Boyandin, 2013; Boyandin et al., 2011; Rey et al., 2020; von Landesberger et al., 2016; Zhao et al., 2008). While handling of space and time dimensions simultaneously is possible, the complexity of time-variant networks (Santoro et al., 2011) pushed researchers to handle the abstraction of time and space dimensions separately. The two alternatives for handling spatial and temporal abstraction separately are: (1) aggregating nodes (spatial units) using flows from all time periods, and then applying temporal clustering, and (2) performing temporal clustering to reduce the number of time steps, and then measuring the similarity of flows between different temporal clusters, and analyzing the changes in spatial patterns over time. Regionalization and clustering methods reduce visual cluttering and generate visually enhanced flow maps with decreased number of flows and regions (or nodes). Grouping of spatial units (nodes) into regions (communities) can also be done for each time period separately (Gao et al., 2013). However, such an approach requires determining cluster correspondence over time and thus makes it difficult to identify changes between time periods. Therefore, previous studies applied regionalization or

clustering to the entire data set to derive groups or regions to be used for the entire temporal extent of time-series flow data (Andrienko et al., 2017; von Landesberger et al., 2016). Moreover, spatial unit-based aggregation methods suffer from the modifiable areal unit problem (MAUP). Issues of spatial scale, aggregation and zoning therefore may result in different flow patterns (Zhu et al., 2019).

To capture change in time, Andrienko et al. (2017) first applied a temporal abstraction by clustering of time intervals based on the similarity of flows. To reduce the number of locations in flow data, Andrienko et al. (2017) clustered locations' flows with a common origin or a common destination by direction and distance ranges of flows, and visualized the mean direction and distances for each location using glyphs. While this approach eliminates the cluttering problem of overlapping lines in flow maps, the connections between locations become difficult to perceive. Von Landesberger et al. (2016) clustered temporal snapshots of flows, so called spatial situations, based on their similarity in different time periods. By computing the mean and median of flows in different time periods, von Landesberger et al. (2016) summarized the spatial situations into clusters of time periods, and thus, decreased the number of time periods to be compared. Our methodology is similar to von Landesberger et al. (2016) in that we evaluate the similarity of flows between time periods to determine the optimal partitioning of time.

Perceptual issues in network visualization

Flow maps require map readers to focus on high-level holistic and geographic tasks such as the comprehension of the dominant flow directions, changing magnitude of flows, spatial communities (strongly connected nodes) and structural changes in geographic flows (Koylu & Guo, 2017). Graph visualization methods can alter the locations of nodes to enhance the readability of graphs and the temporal changes in graph structures. However, nodes are placed in fixed geographic locations and moving the nodes around to enhance readability is not possible for OD flow maps. Several visual variables are used for communicating flow direction, magnitude, and clustering. Arrows and half-arrows have been found to be effective for communicating directional patterns on flow maps (Jenny et al., 2018; Koylu & Guo, 2017). Both line thickness and color gradient have been found to be effective for communicating flow magnitudes (Dong et al., 2018; Jenny et al., 2018; Koylu & Guo, 2017).

Perceptual and cognitive issues in understanding patterns and changes in time-series networks have been extensively studied in dynamic graph visualization literature (Beck et al., 2017; Federico et al., 2011; Federico et al., 2016; Moody et al., 2005; Shi et al., 2011). Archambault et al. (2010) compared animation and small multiples (discrete time-series variable) and found that small multiples produced faster performance than animation whereas animation produced higher accuracy in completing low-level graph comprehension tasks. In contrast, Falkowski et al. (2006) evaluated high level tasks such as identifying and visualizing the evolution of communities (clusters) using a controlled animation method with a time slider. Controlled animation was found to be effective in capturing changes in the evolution of subgroups in communities with a high fluctuation of nodes, i.e., members changing communities (Falkowski et al., 2006). In addition, Groh et al. (2009) adopted Hagerstrand's space-time cube to identify the evolution of node characteristics such as central and prominent nodes as well as the network structures. Findings of user studies that evaluated animated choropleth maps

found that map readers had difficulty detecting changes in animated choropleth maps (Fish et al., 2011). In the light of these findings and the visual clutter generated by flow lines and nodes, we introduce a new methodology to produce time-series flow maps of migration using family trees.

Data

The data set we use in this study is drawn from a larger set of data which contains about 80 million individuals who were born between 1630 and 1930, many of whom were European settlers (Koylu et al., 2020). The U.S. territory and boundaries among the states changed substantially over time. The most common changes were subdivision of larger units into smaller ones as the West was settled. We use 1789, when the first U.S. Congress met and declared that the constitution was in effect, as the starting year of our study period. We end the study period with 1924 based on the Immigration Act of 1924 which limited the number of immigrants who could enter the U.S. by creating quotas based upon national origins. We use historical borders of states and territories as our spatial units using decennial censuses between 1800 and 1930. The number of states (nodes) ranges from 12 in 1789 to 48 in 1924, which is few enough to produce flow maps without having to cluster origins and destinations. This data set allows extraction of migration flows for a period of 135 years, and thus it is ideal for studying long-term changes in migration flows.

Representativeness is one of the limitations of crowd-sourced family trees for studying historical populations. A systematic evaluation of representativeness might be possible if one could link individual family tree records with records from multiple censuses. But such evaluation would still ignore populations such as Native Americans who are not present or underrepresented in the census. And it would have to contend with criteria for linkage between the multiple censuses and between the census and the trees. Attempts to link multiple censuses usually end up with a very small sample, especially of immigrant and non-white groups. To overcome these challenges, we compared the aggregate statistics of family trees to the population statistics of the 1880 Census in our previous work (Koylu et al., 2020). We found that the family trees were biased towards white people, farmers, and native-born (born in the U.S.) Americans compared to the 1880 Census. Also, foreign born and their children, females, and younger people were not represented in trees as much as they were in the census. Race is not available in family tree data, and certain segments of the population are underrepresented in the family trees. For example, Native-Americans and Black population across the country, the Mexican population in the southwest and the Cajun population in Louisiana are underrepresented in the trees. These missing segments of the population is a limitation of family trees as well as census data, which Price et al. (2021) revealed in their study. However, the data are representative of the native-born white population and their migration patterns (Koylu et al., 2020). For further information about the representativeness and data quality of the family tree data we use in this study, please refer to Koylu *et al.* (2020).

Methodology

Our methodology consists of six main steps (Figure 1). First, we extract and date birth events and locations from the family tree data. We then identify family migration using

the child-ladder approach which traces the changes in birthplaces of consecutive children in each individual family. Next, we develop a time-series measure of migration rate and then evaluate a set of algorithmic and domain-based partitioning methods. We then evaluate the changes in migration rate as well as similarity of flows between time periods for each temporal partition to identify an optimal partition. Finally, we construct a time-series of migration networks using a the doubly constrained gravity model and temporal normalization to account for geographic distance, population size, and length differences in time periods. We then generate the time-series flow maps for the optimal periodization to visualize the spatial and temporal patterns of migration.

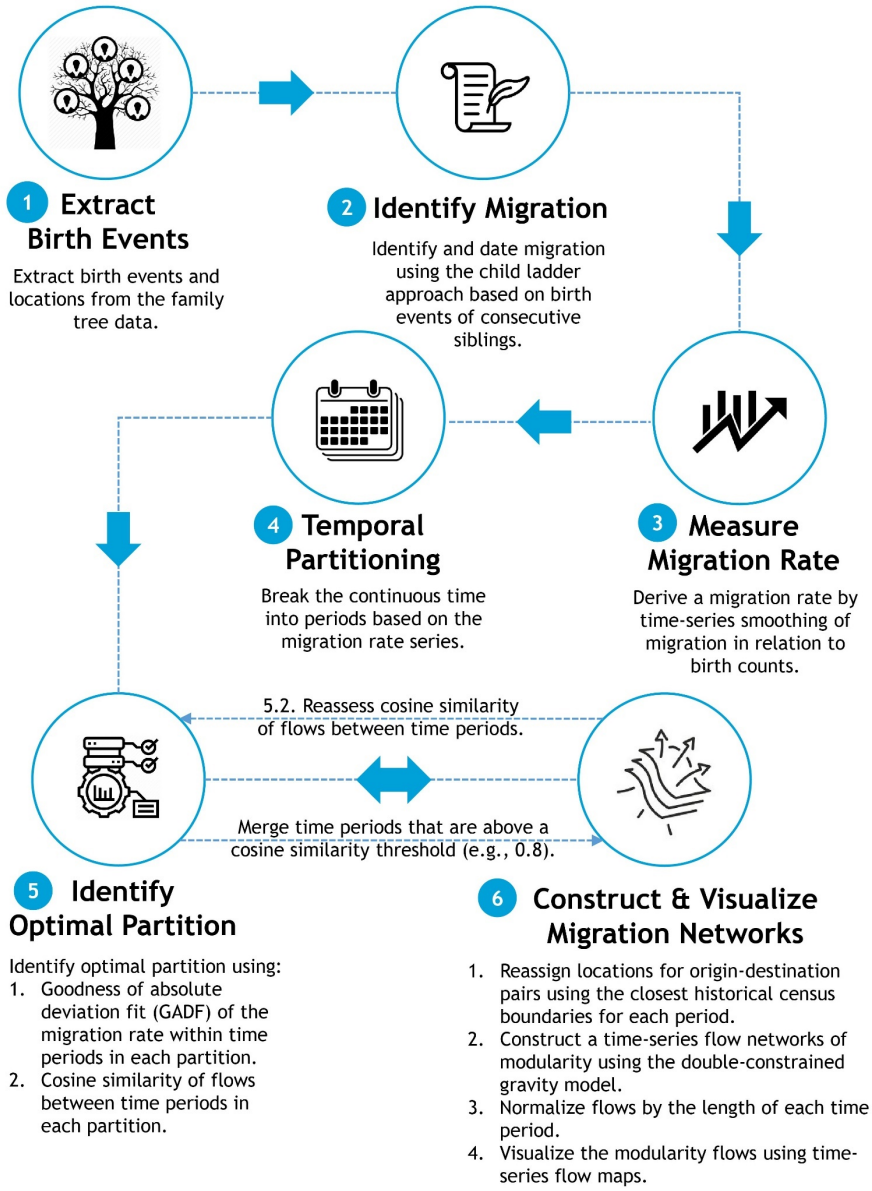


Figure 1. Methodology for measuring and mapping long-term changes in migration flows using population-scale family tree data.

Identifying birth and migration events

There is evidence that larger families moved farther and more often than smaller families both to secure land for their sons and because having several sons reduced the

cost of labor to clear land near the frontier (Adams & Kasakoff, 1984). Moves were frequently undertaken when the father was in his 40's or 50's to maximize family labor (Adams & Kasakoff, 1984). There are two major strategies for extracting migration from birth events of a family: parent-child and child-ladder approaches. The parent-child approach models migration as a move from the birthplace of each parent to the birthplace of each child, and to minimize the effect of family size we have counted children of the same sex and birthplace in the same family just once (Koylu et al., 2020; Koylu & Kasakoff, 2020). The child-ladder approach extracts migration using changes in birthplaces of consecutive siblings in a family (Lathrop, 1948). For example, if there is no change of birthplace between the first child and the second child, then there is no migration. However, if the birthplace of the third child is different than the birthplace of the second, then a migration flow is generated between the second child's birthplace and the third child's birthplace. In this study, we employ the child-ladder approach, and date the move based on the mid-point or average of birth years between the two consecutive siblings with different birthplaces. Throughout the 135 years of our study period between 1789 and 1924, state boundaries and territories substantially change. We use historical census boundaries between 1800 and 1930 to determine locations based on birthyears. We use the closest next census (i.e., the census after the birthyear) and the associated boundaries to locate birthplaces to the states or territories. For example, we locate birthplaces for a migration event in 1792 using the 1800 census boundaries, while we use the 1860 census boundaries to locate birthplaces for an event in 1856.

There are many advantages of the child-ladder approach. First, it allows capturing the moves of a family in a sequence of locations. In contrast, the parent-child approach counts each child born in a different place as a move whose origin is the birthplace of the parent instead of the last location where the family lived. Therefore, the sequence of moves for a family is not captured. Second, the time gap (years) in between a parent's and a child's birth is very large and increases the uncertainty of estimating the migration year. The time gaps between two subsequent children are much smaller. As a result, dating of the moves using the child-ladder approach is more precise. For much of this period children were born approximately two years apart. Third, the parent-child approach double counts moves due to considering both parents' birthplaces and children's birthplaces even if children of the same sex and birthplace can be counted once for reducing such effect. A major limitation of both approaches is that they favor families with more children, families which probably also were more apt to move. Neither approach captures the moves of single individuals or those who did not have children. However, the U.S. population at the time was expanding greatly, and childless individuals were much less common than they were in Europe, probably less than 10% of the population (Hacker, 2016). We also lack information about the portion of the life cycle before and after childbearing. The child-ladder method results in a much smaller set of data than the parent-child method because it is focused on extracting family moves rather than individual moves. However, it is also possible to extract the number of individuals moving with the child-ladder approach by simply adding up living parents and siblings who do not have a death date before the move year.

Measuring the migration rate and temporal partitioning (discretization)

Migration is often compared with a population base to evaluate its relative impact on population redistribution at a global or national scale. Analyzing temporal data over a large span requires partitioning (discretization) of time series, which transforms the

continuous data into discrete time intervals. Partitioning reduces the dimensionality, and the effect of outliers and errors, and makes the knowledge extracted from the discrete intervals easy to understand. However, the choice of how to partition the data, which is a time-varying network, requires significant considerations and experiments to avoid generating anomalies or false patterns that do not naturally exist in the data.

To partition the migration data into periods, we use a migration rate, which we calculate by dividing the family migration counts by the total number of births for a given year. In this section, we first explain why and how we smooth the migration rate over time. Next, we introduce alternative partitioning methods, and our methodology for evaluating the outcome of these methods and identifying the optimal partition.

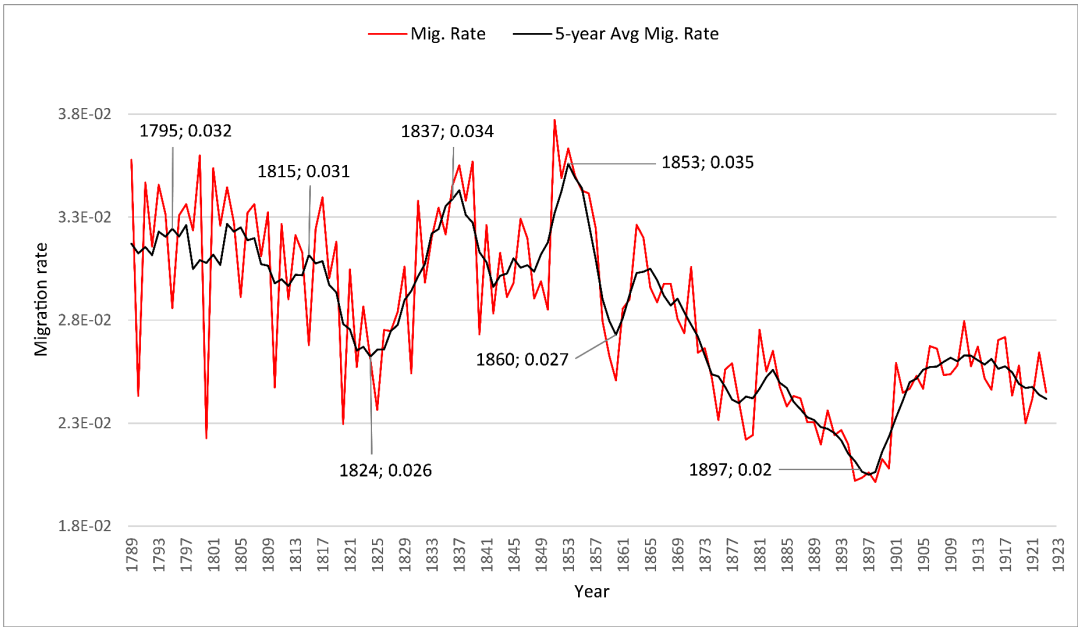
Develop a time-series migration rate

We perform a simple moving average smoothing for the time-series migration and birth data to address the uncertainty in birth years and dating of migration and produce a more robust estimation for the migration rate:

$$mR_t = \frac{\sum_{j=-n}^{n} m_{t+j}}{\sum_{j=-n}^{n} b_{t+j}}$$

mR_t is migration rate for time period *t* that is derived by dividing the total number of family migration (m_t) by the number of births (b_t) in each year using a moving average window. m_t is the migration count at time t which is smoothed by a moving average of order (length in years): $w = 2n + 1$. For example, for a five-year moving average window, *w* equals to 5. Therefore, the number of periods *(n)* to be included in smoothing before and after the center (estimation) period of *t* is 2. The estimates of *m* and *b* at time *t* are obtained by averaging the time series within *n* periods of *t*. Because mR_t is a division of two moving average values of migration and birth counts, the division of the sum of weights by w (i.e., *1 / w*) gets cancelled out, thus, is not included in the formula.

Our goal is to smooth the data as little as possible to remove large fluctuations and keep the temporal distribution and trends as similar as possible to the original data. We first analyze the distribution of the number of years between consecutive siblings (see Appendices Figure A.1). The mean migration cycle is 4.08 years, and the median is 3 years. About 58% of all births are within 3 years, 77 % within 5 years, and 94% within 10 years. To address the uncertainty related to birth years and mid-point dating of migration, we perform experiments on smoothing windows based on odd number window lengths from 3 to 13. We choose 5-year moving window that is long enough to capture the mean migration cycle (4.08) and accounts for 77% of birth year differences among consecutive siblings. Figure 2 illustrates the original yearly migration rate (red) and 5-year moving average migration rate (black). The rate starts high but then falls before rising to twin peaks at the middle of the 19th century from which it declines sharply to a low in 1897. It rebounds a bit but continues at a lower rate from the end of the 19th century to 1924 when the analysis ends.



Temporal partitioning methods

Next, we partition the time based on the migration rate using four different methods: historical periods, transportation periods, and periods derived from two data-driven methods: equal (fixed) interval and temporal natural breaks. All partitions start with 1789 Constitution of the U.S. and end with the Immigration Act of 1924. We generate from 4 to 32 consecutive temporal periods for both equal interval and temporal natural breaks methods.

Historical and transportation periods: We partition the time into six historical periods and five transportation periods (Table 1). The first three periods are based on standard eras and breakpoints in American history such as the Early Republic and Pre-transportation Revolution (1776-1823), Transportation Revolution: 1824-1840, and Pre-civil War: 1841-1860. We split the post transportation revolution period into a pre-war and post-war period based on the fact that the Republican Congress passed the Homestead Act during the war and funded the Transcontinental Railroad, which had a significant impact on internal migration.

Table 1: Historical and transportation periods

Historical periods		Transportation periods	
1776-1823	Early Republic and Pre-transportation Revolution		
1824-1840	Transportation Revolution		
1841-1860	Pre-civil War		
1861-1865	Civil War	1861-1889	Civil War, Homestead Act, Transcontinental Railroad and the closing the Frontier

1866-1900	Post Bellum and Rise of Industrial America	1890-1924	Post Bellum to Progressive Era: 1890-1924.
1900-1924	Age of Immigration / Progressive Era		

Equal interval (fixed) periods: We partition the time into equal intervals of 5 to 20 years (i.e., 5, 6, …19, 20), and 25, 35, 40, 45 and 50 years of length. We obtain a total of 22 partitions. For example, the time periods for a five-year period length are 1789-1794, 1794-1799…, and 1921-1924. The number of time periods varies for each fixed length periodization. The five-year length produces 27 periods, 10-year produces 14 periods, 15-year produces 9 periods, and 20-year produces 7 periods.

Temporal natural breaks periods: We partition the time using the natural-Jenks classification, which captures natural breaks of time periods for the migration rate. The natural breaks algorithm maximizes the similarity of values (the migration rate) within each period. Different from the conventional natural breaks classification that first orders numerical values, temporal natural breaks use the data in its original order of time (years). Natural breaks partitioning also requires the number of time periods to be determined. We obtain a total of 29 partitions with the total number of time periods varying between 4 and 32.

Identifying an optimal temporal partition

We evaluate two measures to identify an optimal temporal partition: (1) goodness of absolute deviation fit (GADF) that measures the homogeneity of the migration rate within time periods, and (2) cosine similarity (CS) that measures the similarity of flows between time periods in a partition. While our goal is to maximize the similarity of the migration rate within periods using GADF, we aim to maximize the difference between flows in consecutive time periods using CS. GADF value of a partition is calculated as below (Slocum et al., 2009):

$$\mathrm{GADF} = 1 - (\mathrm{ADCM}/\mathrm{ADAM})$$

ADCM is the sum of absolute deviations about class medians for a particular number of classes; and ADAM is the sum of absolute deviations about the median for the entire data set. GADF ranges from 0 to 1, with 0 representing the lowest accuracy (one period) and 1 representing the highest accuracy (t periods for t-size temporal records).

In addition to capturing changes in the migration rate using the GADF evaluation, capturing how local patterns of flows change over time is also important for identifying an optimal temporal partition. To capture the correlation and changes in spatial flows (OD volumes), we calculate cosine similarity of flow matrices between consecutive periods in each partition. Cosine similarity for a pair of networks is calculated by taking the inner product space that measures the cosine of the angle between two non-zero vectors of flow values that form each network. In our case, each time series network consists of a non-zero vector that is formed by a list of origin-destination (OD) pairs and their corresponding flow values.

Figure 3.a illustrates GADF values that indicate the homogeneity of the migration rate within time periods in each partition, whereas Figure 3.b illustrates median CS values that indicate the similarity of flows between time periods in each partition. We expect GADF values to be optimally as high as possible, while we expect CS value for a partition to be as low as possible. Both Figure 3.a and 3.b plot the number of periods against GADF and CS values for each partitioning method. Two common approaches for selecting the appropriate number of periods for GADF are to choose a threshold such as 0.8 or to identify a point at which the curve begins to flatten out. A flattening at a point would indicate that a larger number of periods would not result in the reduction of classification error, which can be identified by a decreasing slope of the GADF value. The temporal natural breaks method outperforms other methods regardless of the number of periods based on GADF evaluation. This is not surprising as the natural breaks algorithm optimizes the GADF values by minimizing the within-class similarity of the partitions. In Figure 3.a, the seven-period partition is the first periodization in which the degree of slope for GADF decreases. On the other hand, in Figure 3.b, we aim to identify a break in which the increasing trend reaches a plateau or is reversed. Fixed-length periods produce comparable results to temporal natural breaks for median CS values for partitions with small number of periods. The two cases with a break in the increasing trend exist in natural breaks periods with seven and 11 periods. We choose temporal natural breaks with seven periods as an optimal choice because of the agreement between GADF and CS measures. Figure 4 illustrates the migration rate with the temporal natural breaks in our optimal partition with seven periods. Table 2 illustrates CS values between all periods in our optimal partition with seven periods. Table 2 highlights that the flows are more similar between consecutive periods with a median of 0.73.

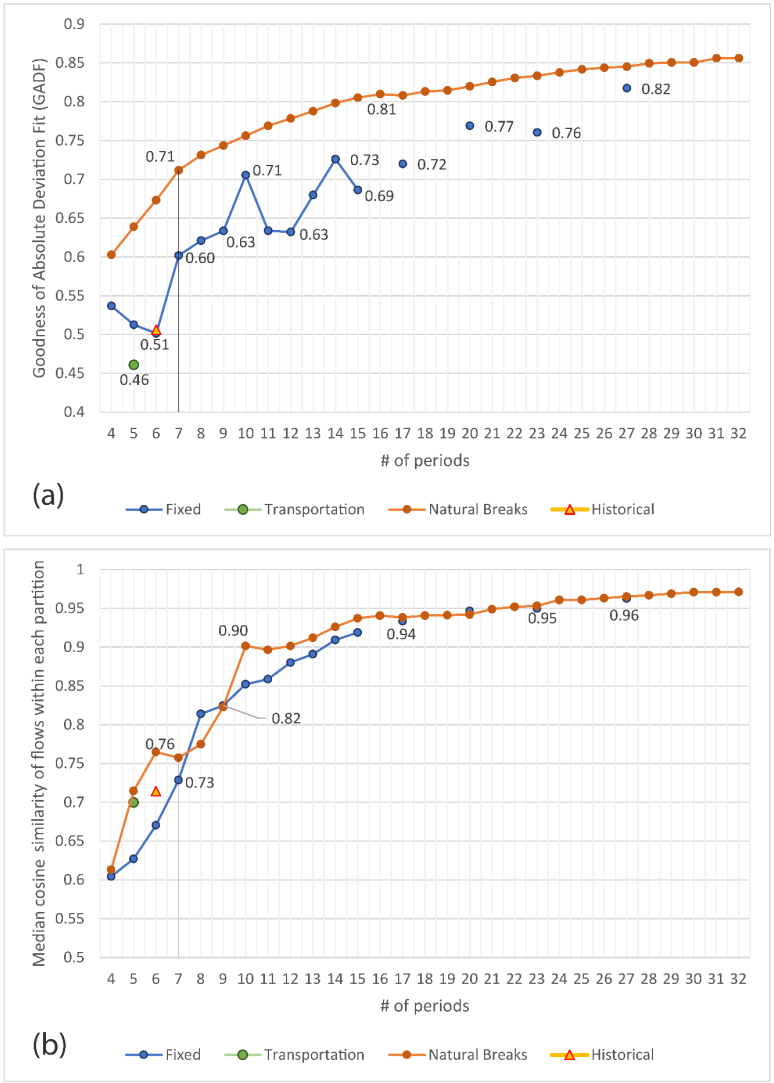


Figure 3: (a) Goodness of absolute deviation fit (GADF) of the migration rate that indicates the homogeneity of migration rates within each period in a partition. (b) Median cosine similarity (CS) of flows that indicates how similar spatial patterns of flows between time periods in each partition. A larger value of GADF and a lower value of median CS is needed to identify an optimal partition. We chose temporal natural breaks with seven periods as our optimal periodization.

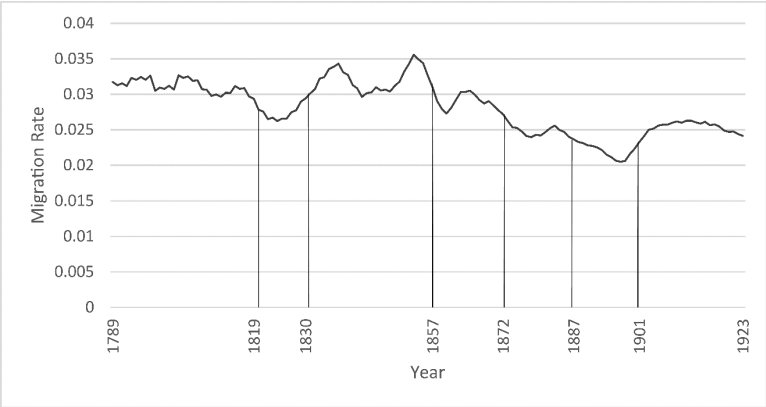


Figure 4: Temporal natural breaks with the seven periods.

Table 2: Cosine similarity of flows between time periods of the optimal partition.

	1789-1819	**1819-1830**	**1830-1857**	**1857-1872**	**1872-1887**	**1887-1901**
1819-1830	0.76					
1830-1857	0.47	0.77				
1857-1872	0.23	0.40	0.66			
1872-1887	0.19	0.30	0.52	0.75		
1887-1901	0.18	0.25	0.38	0.52	0.74	
1887-1924	0.20	0.26	0.33	0.43	0.60	0.91

Construct and visualize migration networks

An important issue with temporal partitioning is the necessity for the re-assignment of the locations, i.e., origins and destinations that come from birthplaces of consecutive siblings, after merging years into periods. For example, the migration event in 1792 is located based on the boundaries of the 1800 census for calculating the migration rate. However, when we generate time periods such as from 1789 to 1819, we aggregate the migration events between the start and end of the period. We assign birthplaces to states and territories using the closest census after the migration event, which is the 1820 census for the period that ends in 1819. Consequently, the same migration event in 1792 that is located based on 1800 census boundaries now gets located based on 1820 boundaries due to the temporal periodization and aggregation for the 1789-1819 period.

Modularity normalization

We generate a series of migration networks using the temporal natural breaks with the seven periods. We transform the raw flows in each time period into modularity flows (Newman, 2006) using the doubly constrained gravity model (Roy & Thill, 2004) to account for the effect of geographic proximity and the ability of states to generate migration flows. We calculate the modularity between a pair of states (i and j) using the formula:

$$\text{Modularity (i, j)} = \text{Observed Flows } (F_{ij}) - \text{Expected Flows } (E_{ij})$$

where F_{ij} is the number of observed flows and E_{ij} is the expected number of flows from state i to state j. After subtracting expected flows from observed flows, we derive a measure of modularity for each origin-destination pair. While a positive modularity value indicates that the observed flow volume is above the expectation, a negative value indicates that the observed flow volume is less than the expectation.

We employ the doubly constrained gravity model that constrains both origins and destinations by enforcing two rules: (1) the sum of expected flows from an origin is equal to the observed, and (2) the sum of expected flows to a destination is equal to the observed volume of flows to that destination.

$$E_{ij} = \mathrm{A}_i * \mathrm{O}_i * \mathrm{B}_j * D_j * \mathrm{D}_{ij}^{-beta}$$

where O_i and D_i are the total of out-flows and in-flows of state i, respectively. A_i and B_i are the balance factors that are calculated by the following iteration. While the distance decay function is square and uniform for all states, each node (state) has a different set of parameters.

$$A_i = 1/\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\mathrm{B}_j\mathrm{D}_j * \mathrm{D}_{ij}^{beta}\right)$$

$$B_i = 1/\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\mathrm{A}_j\mathrm{O}_j * \mathrm{D}_{ij}^{beta}\right)$$

Merging periods with similar flow patterns

After the modularity transformation, one can decide whether to merge periods by re-evaluating the similarity of flows between consecutive periods. Table 3 illustrates the cosine similarity of flows after the gravity-based modularity transformation. Median cosine similarity between consecutive periods is about 0.71. The most similar time periods are 1887-1901 and 1901-1924 with a value of 0.84. Based on this result, we do not merge the periods in this study.

Table 3: Cosine similarity of the gravity-normalized modularity flows between time periods of the optimal partition.

	1789-1819	**1819-1830**	**1830-1857**	**1857-1872**	**1872-1887**	**1887-1901**
1819-1830	0.78					
1830-1857	0.46	0.66				
1857-1872	0.21	0.28	0.64			
1872-1887	0.14	0.16	0.32	0.61		
1887-1901	0.15	0.14	0.23	0.40	0.75	
1887-1924	0.20	0.17	0.18	0.29	0.53	0.84

Temporal normalization

The seven time periods resulting from optimization vary substantially. In order to make maps of flows during these periods comparable, we divide the gross volume of flows (i.e., illustrated by point symbols) and the modularity values (i.e., illustrated by the flow symbols) by the number of years in each period. Consequently, our maps illustrate yearly average gross volume and modularity flows.

Results

Figures 5, 6, 7 and 8 illustrate the time-series flow maps of family migration during the seven optimal periods between 1789 and 1924, which we produced using FlowMapper.org (Koylu, Tian, & Windsor, 2021). Flows are curvy at the origin and straight at the destination end with a half-arrowhead to enhance the readability of flow lines. Flow symbols illustrate the yearly average modularity flows between states that highlight migration that are above our expectation and thus incorporate a correction for "gravity effect". Flow line thickness is classified into three classes 1-50, 50-100, and 100 to 150. Point symbols are proportional to yearly average gross volume of flow: the number of families that moved from and into each state. The choropleth map illustrates the migration efficiency, which is calculated for each state using the formula below.

$$Migration\ efficiency = (InMigrants - OutMigrants)\ /\ (InMigrants + OutMigrants)$$

Unsurprisingly, the dominant pattern is westward migration. The losing area in the East gets larger over time and the gaining area moves westward. There is always a "frontier", i.e., states and territories in the West gaining population through migration. The migration from the East to the West is largely confined to longitudinal bands (east-west direction) which would be clearer if we could use counties instead of states. There were three bands in the south: Virginia (VA) to Kentucky (KY), North Carolina (NC) to Tennessee (TN) and migration along the entire lower South (Figure 5.a and b). The three bands in the South continue until 1857 when the migration rate drops in the South and becomes similar to more northern states on the East Coast (Figure 6). From Figure 6, we can also see the Gold Rush to California and Oregon even though that was said to be mostly single men (Hurtado, 1999).

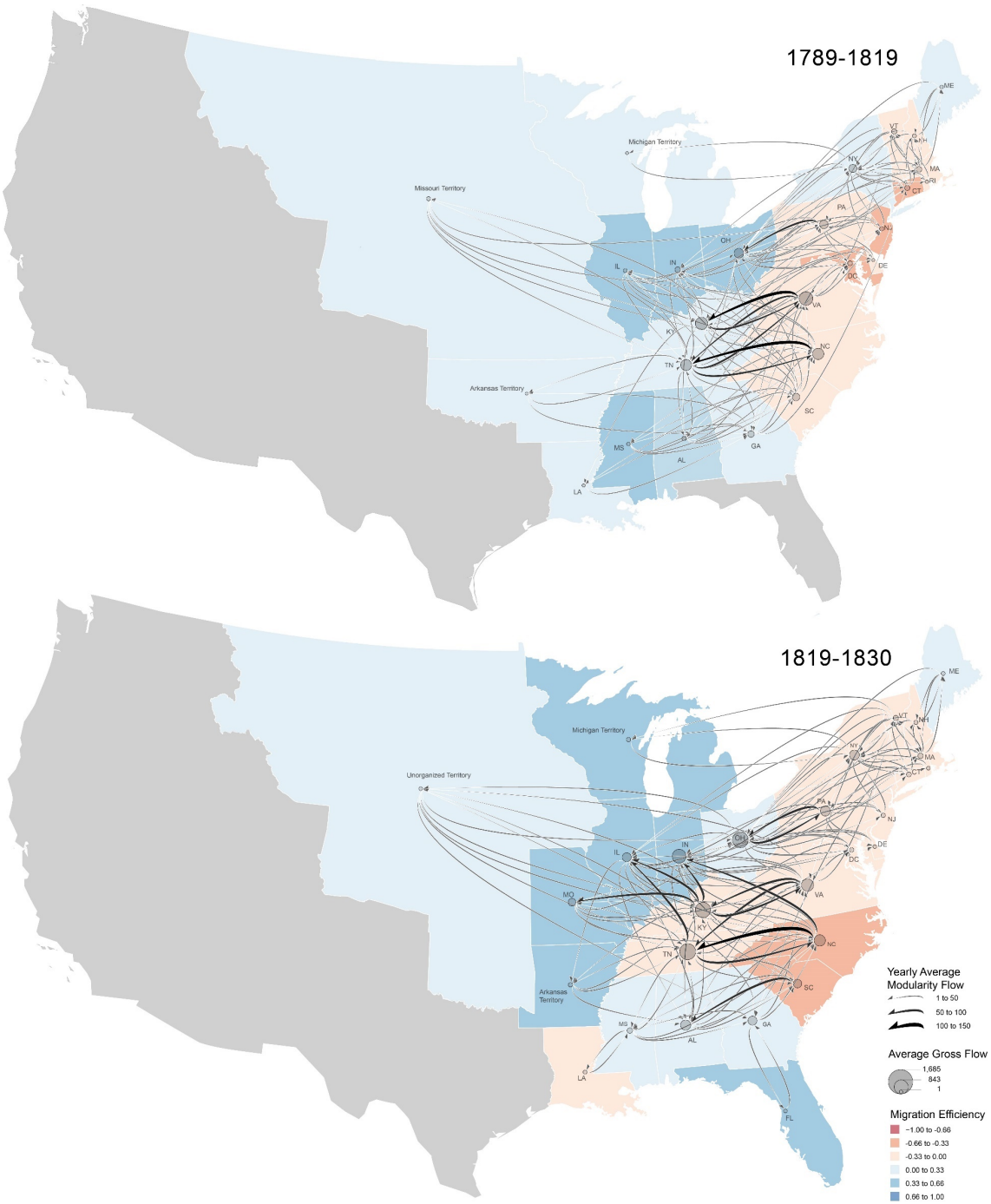


Figure 5: Family migration in (a) 1789-1819 and (b) 1819-1830. Flows illustrate yearly average modularity flow between pairs of states; nodes illustrate yearly average gross volume of flows per state; and the choropleth map illustrates the migration efficiency. We subtract the expected volume of flows calculated based on the doubly constrained gravity model from the observed volume of flows to derive modularity. We normalize modularity and gross volume of flows by the total number of years in each period and derive yearly average modularity and gross volumes so that maps from multiple time periods can be compared.

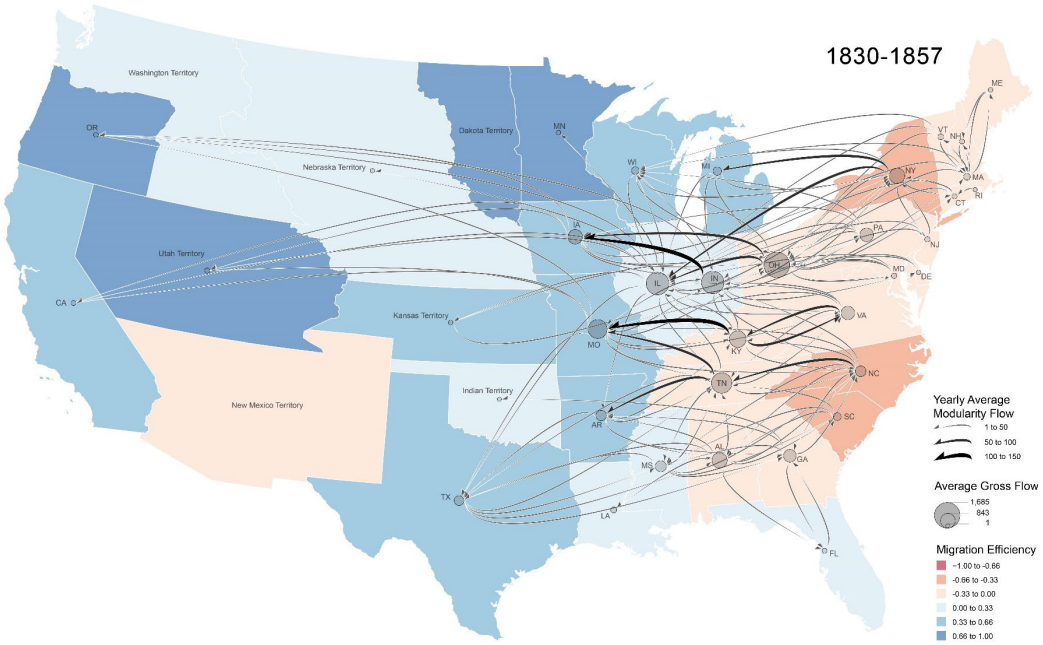


Figure 6: Family migration in 1830-1857. Flows illustrate yearly average flow above expectation (modularity) between pairs of states; nodes illustrate yearly average gross volume of flows per state; and the choropleth map illustrates the migration efficiency.

The changing pattern of point symbols show the amount of migration, the gross flow, which moved westward, especially after 1872 (Figure 7.a and b). Utah is clearly a hearth after 1872, sending migrants to all the adjacent states (Figure 7.b). The frontier is fed from the previous frontier, therefore, long distance migration from the East slowed over time but some Eastern states start to gain again, most likely a result of migration into cities on the East Coast such as Boston and New York. The losses from rural areas of New York outweigh New York City's gains, however, and the circular migration between adjacent states in the East increases over time. However, it is important to point out that these maps do not include immigrants from other countries who largely settled in cities in the Eastern part of the country.

Beginning in 1887 and through 1901, the southern and the middle bands in the South both send migrants to Texas (Figure 8.a). If we were mapping at a higher resolution such as counties, we might see flows going to different parts of the state. The maps grow lighter over time and the gross flows in and out of each state shrink, due to decreasing rates of migration which spreads from East to West until during the last period there is only one state showing large gains: California (Figure 8b). There are no points as large as they were during the period of the twin peaks, 1830 to 1857 Migration into Oklahoma around the oil boom is clearly visible (Figure 8a). The final period also shows some very long-distance flows, i.e., from New York (NY) and Pennsylvania (PA) to California (CA), and from Michigan (MI) and Wisconsin (WI) to Oregon (OR) as well as long distance flows from the West to the East such as from California (CA) to

Texas (TX) (Figure 8.b). These patterns may be an outcome of the migration between large cities, a later pattern quite important today.

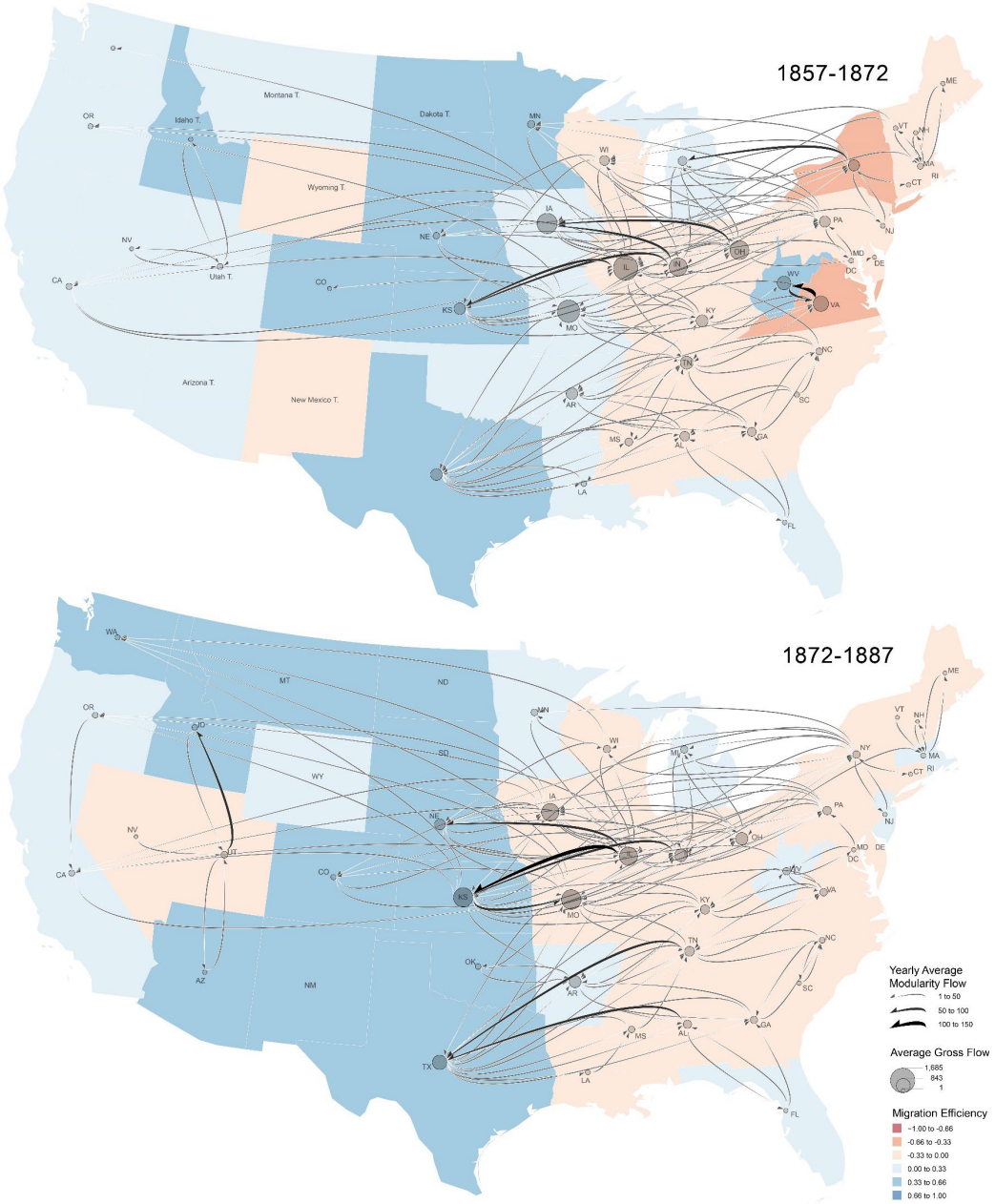


Figure 7: Family migration in (a) 1857-1872 and (b) 1872-1887. Flows illustrate yearly average flow above expectation (modularity) between pairs of states; nodes illustrate yearly average gross volume of flows per state; and the choropleth map illustrates the migration efficiency.

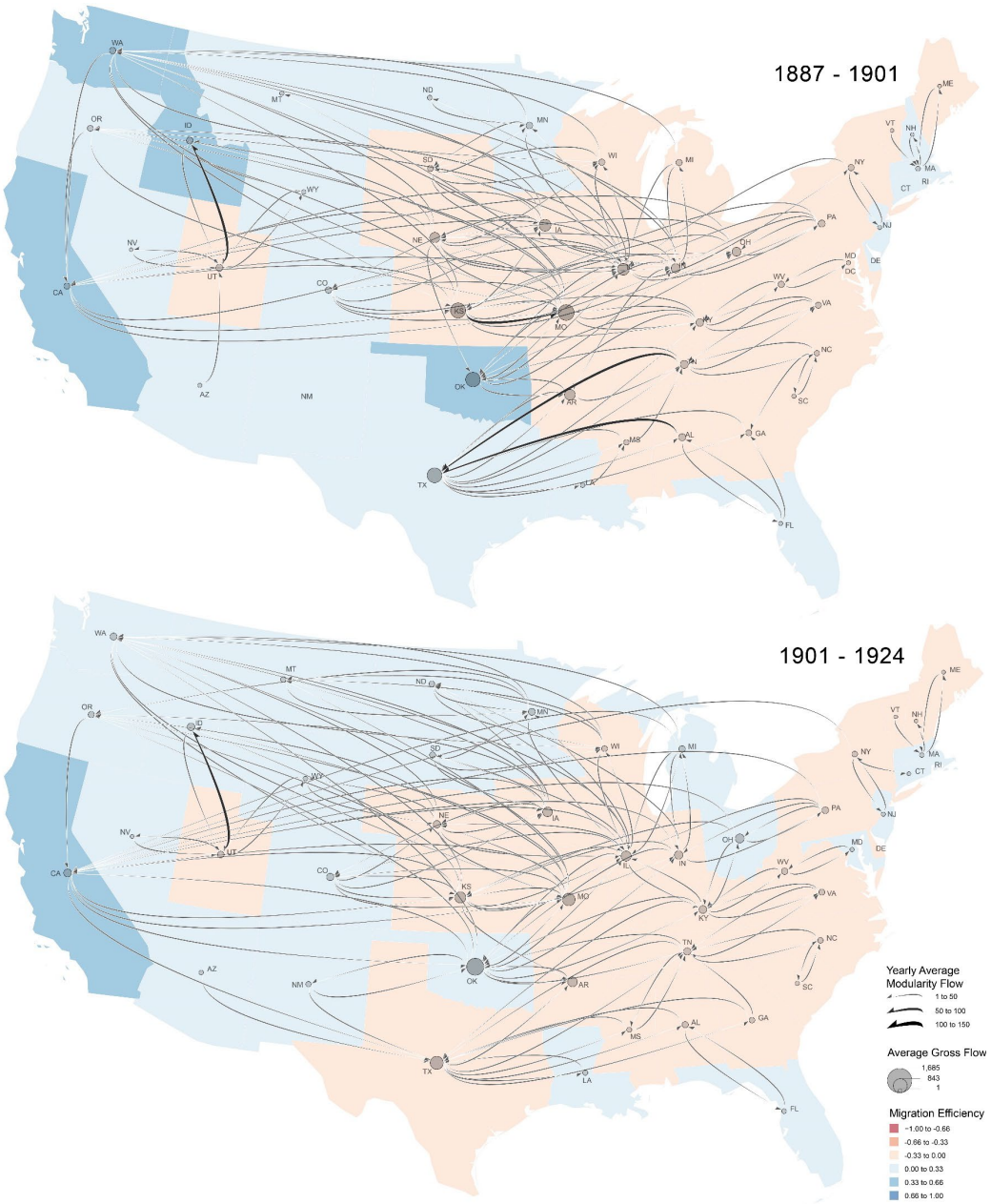


Figure 8: Family migration in (a) 1872-1887 and (b) 1901-1924. Flows illustrate yearly average flow above expectation (modularity) between pairs of states; nodes illustrate yearly average gross volume of flows per state; and the choropleth map illustrates the migration efficiency.

Discussion and Conclusion

This study represents a first attempt to measure and map long-term changes in migration flows using a population-scale family tree data set. It is only possible because of the data we have been able to use, which stretches over a long period of time, an important example of colonization which unfolded in a similar fashion in many different parts of the globe during the 17th to 20th centuries.

To the best of our knowledge, these maps are the first maps to systematically show how the internal migration flows within the U.S. changed over a long period of time (i.e., 135 years). Previous maps showed only the movement of the center of the U.S. population westward (Census.gov) or used dots or density maps to show the changing distribution of the population but did not show the specific flows that led to these changes. Also, we attempt to identify the actual turning points in these flow patterns using continuous data instead of depicting static snapshots that census could provide.

Flows have several properties which change over time: volume, direction, spatial and topological (connectivity) patterns. We evaluate domain-based and data-driven periodization considering both the similarity of the migration rate within periods, and the similarity of spatial flow patterns between periods. We identify the temporal breaks with seven periods as the optimal partition in both evaluations. Whether the result of the two evaluations is a coincidence or structural would require further analysis but it does make sense that changes in the migration rate would be associated with changes in origins and destinations, especially if the dominant process is colonization.

The flows we map are composed of family units, but there were others travelling these same paths: single individuals, both men and women. Their flows would likely differ from each other and from the families we are studying here. In addition, the migration we study from the population-scale family tree is dominated by the patterns of the earliest European settlers (Koylu et al., 2020). Later immigrant groups are much less represented, and it is likely that there are very few Native Americans, Blacks, Hispanics in these family trees. Some of the forces, such as transportation and economic opportunities, which channeled family flows would also channel individual flows, likewise these would also have affected the migration of all the ethnic groups, but the volume of the flows would be different which could lead to different flows being highlighted in the maps. This is a topic for further research. Although the sample is most representative of the native-born white population, it is very large with many examples of later immigrant groups. The census can be used to check whether the smaller samples we have from these groups are representative of their locations. Kandt, van Dijk and Longley (2020) extracted migration flows using surnames in individual level historic census data in Great Britain. Their study shows great promise for the use census data combined with family records for studying ethnic origins and intergenerational change in local populations across space and time.

The ability to detect patterns depends upon the precision of dates for migration. In previous work we used a very imprecise time window to determine dates: birth places of parents and their children (Koylu & Kasakoff, 2020), usually between 20 and 40 years. But in this article, we use the child-ladder approach to compare different temporal partitions with reduced the margin of error for dating migration. This also allows us to see how certain historical events of shorter duration affected migration.

The flows are so complex that it is not possible to say that a particular periodization captures change better than any other. It depends on the story one wants to tell. In this case, the story would have to include westward expansion. To tell that story visually, one might want to eliminate the flows to adjacent states unless they are above a certain threshold. We use the gravity model to adjust for this. The flows to adjacent states, which become more important in the East over time, may represent neighborhood moves that just happen to be across state borders or moves to urban or industrial centers drawing families from nearby states. Analysis at a smaller spatial scale than the state could clarify this and reveal a great deal more about these flows.

A limitation of this work that needs to be addressed is the effect of the decline in fertility upon our results. Our maps highlight flows in areas that were just being settled because our ability to detect moves depends upon the birth of children. There are several remaining questions. How much of the slowing of migration in the East over time was due to the fertility decline, which would lead to fewer family moves being detected, and how much was due to changing migration patterns on the part of the families in the settled areas who moved shorter distances and perhaps less frequently than families closer to the frontier? The issue can be addressed by a spatial analysis of the change in fertility over time which can be used to adjust the flows.

Aside from the need for future research in migration and family trees, there is also a need for new cartographic theories and methods for mapping large and complex space-time flow data (Andrienko et al., 2008; Dodge, 2019; Griffin, Robinson, & Roth, 2017; Robinson et al., 2017; Tsou, 2015). In a recent review of the literature in computational and visual movement analytics, Dodge and Noi (2021) argue the need for a generic cartographic framework describing a set of visual principles for mapping movement and guiding the evaluation of movement visualizations in different applications. There is an emerging body of literature on the principles of flow map design and user studies (Jenny et al., 2018; Koylu & Guo, 2017; Yang et al., 2017; Yang et al., 2019). Despite these recent efforts, there is a lack of knowledge about how map readers identify changes in time-series flow maps. Specifically, the effect of interval size, number, and positioning on visual understanding of temporal patterns of flows has not been studied. Such effects could result in important differences whether the flow maps are animated or presented as a series of temporal periods. For example, a sudden change in flow patterns could appear gradual if the time intervals are larger. Characteristics of flows such as volume, direction, and topology interact and influence each other differently for each dataset, layout, and temporal scale in time-series flow networks. How to balance all of these in a single analysis is an issue for much discussion and future research.

Data and codes availability statement

The family tree data used in this study were derived from the following resources available in the public domain: https://home.rootsweb.com between the dates of February and August 2015. Based on the effective date of the Revised Terms and Conditions of Ancestry.com by 25 Jul 2019, we are not able to share the original GEDCOM files published on rootsweb.com. We share the anonymous birth events and locations as a database backup file (at state and territory level) that can be used for extracting the child-ladder migration. We share the database backup file and the medium and final data products at the following data source link: https://doi.org/10.6084/m9.figshare.14602677.v3. We share the entire source code for

our methodology at the following link:
https://doi.org/10.6084/m9.figshare.14602671.v3

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Adams, J. W., & Kasakoff, A. B. (1984). Migration and the family in colonial New England: The view from genealogies. *Journal of Family History, 9*(1), 24-43. doi:10.1177/036319908400900102

Adrienko, N., & Adrienko, G. (2011). Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics, 17*(2), 205-219. doi:10.1109/TVCG.2010.44

Andrienko, G., Andrienko, N., Dykes, J., Fabrikant, S. I., & Wachowicz, M. (2008). Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. *Inf Visualization, 7*(3-4), 173-180. doi:10.1057/ivs.2008.23

Andrienko, G., Andrienko, N., Fuchs, G., & Wood, J. (2017). Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data. *IEEE transactions on visualization and computer graphics, 23*(9), 2120-2136. doi:10.1109/TVCG.2016.2616404

Archambault, D., Purchase, H., & Pinaud, B. (2010). Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE transactions on visualization and computer graphics, 17*(4), 539-552. doi:10.1109/TVCG.2010.78

Beck, F., Burch, M., Diehl, S., & Weiskopf, D. (2017). *A taxonomy and survey of dynamic graph visualization.* Paper presented at the Computer Graphics Forum.

Beecham, R., & Wood, J. (2014). Characterising group-cycling journeys using interactive graphics. *Transportation Research Part C: Emerging Technologies, 47*, 194-206. doi:10.1016/j.trc.2014.03.007

Bertin, J. (2010). Semiology of graphics: Diagrams, networks, maps. In. Redlands, CA: ESRI Press.

Boyandin, I. (2013). *Visualization of temporal origin-destination data.* University of Fribourg,

Boyandin, I., Bertini, E., Bak, P., & Lalanne, D. (2011). *Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data.* Paper presented at the Computer Graphics Forum.

Dodge, S. (2019). A Data Science Framework for Movement. *Geographical Analysis*. doi:10.1111/gean.12212

Dong, W., Wang, S., Chen, Y., & Meng, L. (2018). Using eye tracking to evaluate the usability of flow maps. *ISPRS International Journal of Geo-Information, 7*(7), 281. doi:10.3390/ijgi7070281

Dorling, D. (1998). Human cartography: when it is good to map. *Environment and Planning A, 30*(2), 277-288.

Falkowski, T., Bartelheimer, J., & Spiliopoulou, M. (2006). *Mining and Visualizing the Evolution of Subgroups in Social Networks*. Paper presented at the Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

Federico, P., Aigner, W., Miksch, S., Windhager, F., & Zenk, L. (2011). *A visual analytics approach to dynamic social networks*. Paper presented at the Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, Graz, Austria.

Federico, P., & Miksch, S. (2016). *Evaluation of two interaction techniques for visualization of dynamic graphs.* Paper presented at the International Symposium on Graph Drawing and Network Visualization.

Fish, C., Goldsberry, K. P., & Battersby, S. (2011). Change blindness in animated choropleth maps: an empirical study. *Cartography and Geographic Information Science, 38*(4), 350-362. doi:10.1559/15230406384350

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS, 17*(3), 463-481. doi:10.1111/tgis.12042

Griffin, A. L., Robinson, A. C., & Roth, R. E. (2017). Envisioning the future of cartographic research. *International Journal of Cartography, 3*(sup1), 1-8. doi:10.1080/23729333.2017.1316466

Groh, G., Hanstein, H., & Wörndl, W. (2009). *Interactively Visualizing Dynamic Social Networks with DySoN.* Paper presented at the Computer-Human Interaction (CHI), Boston.

Guo, D. (2009). Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 1041-1048. doi:10.1109/TVCG.2009.143

Guo, D., & Zhu, X. (2014). Origin-Destination Flow Data Smoothing and Mapping. *Visualization and Computer Graphics, IEEE Transactions on, PP*(99), 1-1. doi:10.1109/TVCG.2014.2346271

Hacker, J. D. (2016). Ready, willing, and able? Impediments to the onset of marital fertility decline in the United States. *Demography, 53*(6), 1657-1692. doi:10.1007/s13524-016-0513-7

Han, E., Carbonetto, P., Curtis, R. E., Wang, Y., Granka, J. M., Byrnes, J., . . . Ball, C. A. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nature Communications, 8*(1), 14238. doi:10.1038/ncomms14238

Holland, S. C., & Plane, D. A. (2001). Methods of mapping migration flow patterns. *Southeastern Geographer, 41*(1), 89-104.

Hollingsworth, T.-H. (1970). *Historical studies of migration.* Paper presented at the Annales de démographie historique.

Holten, D., & van Wijk, J. J. (2009). Force-Directed Edge Bundling for Graph Visualization. *Computer Graphics Forum, 28*(3), 983-990. doi:10.1111/j.1467-8659.2009.01450.x

Hurtado, A. L. (1999). Sex, gender, culture, and a great event: The California gold rush. *Pacific Historical Review, 68*(1), 1-19. doi:10.2307/3641867

Hägerstrand, T. (1976). Geography and the study of interaction between nature and society. *Geoforum, 7*(5-6), 329-334. doi:10.1016/0016-7185(76)90063-4

Jenny, B., Stephen, D. M., Muehlenhaus, I., Marston, B. E., Sharma, R., Zhang, E., & Jenny, H. (2018). Design principles for origin-destination flow maps. *Cartography and Geographic Information Science, 45*(1), 62-75. doi:10.1080/15230406.2016.1262280

Kandt, J., Cheshire, J. A., & Longley, P. A. (2016). Regional surnames and genetic structure in Great Britain. *Transactions of the Institute of British Geographers, 41*(4), 554-569. doi:10.1111/tran.12131

Kandt, J., van Dijk, J., & Longley, P. A. (2020). Family Name Origins and Intergenerational Demographic Change in Great Britain. *Annals of the American Association of Geographers, 110*(6), 1726-1742. doi:10.1080/24694452.2020.1717328

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., . . . Gymrek, M. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science, 360*(6385), 171-175. doi:10.1126/science.aam9309

Kasakoff, A. B. (2019). The Changing Space of Families: A Genealogical Approach. *Social Science History, 43*(1), 1-29. doi:10.1017/ssh.2018.39

Koylu, C., & Guo, D. (2017). Design and evaluation of line symbolizations for origin–destination flow maps. *Information Visualization, 16*(4), 309-331. doi:10.1177/1473871616681375

Koylu, C., Guo, D., Huang, Y., Kasakoff, A., & Grieve, J. (2020). Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S. *International Journal of Geographical Information Science*. doi:10.1080/13658816.2020.1821885

Koylu, C., & Kasakoff, A. (2020). *Mapping temporal trends of parent-child migration from population-scale family trees* Paper presented at the AutoCarto International Research Symposium, World Wide Web.

Koylu, C., Tian, G., & Windsor, M. (2021). FlowMapper.org: A web-based framework for designing origin-destination flow maps. *Journal of Maps*. doi:10.1080/17445647.2021.1996479

Kraak, M. J. (2003). The space-time cube revisited from a geovisualization perspective. *Proceedings of the 21st International Cartographic Conference, 1988-1995*.

Lathrop, B. F. (1948). Migration into East Texas 1835-1860. *The Southwestern Historical Quarterly, 52*(1), 1-31.

Lobben, A. (2003). Classification and application of cartographic animation. *The Professional Geographer, 55*(3), 318-328. doi:10.1111/0033-0124.5503016

Moody, J., McFarland, D., & Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology, 110*(4), 1206-1241.

Nelson, M. A. (2020). The decline of patrilineal kin propinquity in the United States, 1790–1940. *Demographic Research, 43*, 501-532. doi:10.4054/DemRes.2020.43.18

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences, 103*(23), 8577-8582. doi:10.1073/pnas.0601602103

Otterstrom, S. M., & Bunker, B. E. (2013). Genealogy, migration, and the intertwined geographies of personal pasts. *Annals of the Association of American Geographers, 103*(3), 544-569. doi:10.1080/00045608.2012.700607

Pooley, C. G., & Turnbull, J. (1997). Leaving home: the experience of migration from the parental home in Britain since c. 1770. *Journal of Family History, 22*(4), 390-424. doi:10.1177/036319909702200402

Price, J., Buckles, K., Van Leeuwen, J., & Riley, I. (2021). Combining family history and machine learning to link historical records: The Census Tree data set *. *Explorations in Economic History, 80*. doi:10.1016/j.eeh.2021.101391

Rey, S., Han, S. Y., Kang, W., Knaap, E., & Cortes, R. X. (2020). A Visual Analytics System for Space–Time Dynamics of Regional Income Distributions Utilizing Animated Flow Maps and Rank-based Markov Chains. *Geographical Analysis*.

Robinson, A. C., Demšar, U., Moore, A. B., Buckley, A., Jiang, B., Field, K., . . . Sluter, C. R. (2017). Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *International Journal of Cartography, 3*(sup1), 32-60. doi:10.1080/23729333.2016.1278151

Roy, J. R., & Thill, J. C. (2004). Spatial interaction modelling. *Papers in Regional Science, 83*(1), 339-361. doi:10.1007/s10110-003-0189-4

Santoro, N., Quattrociocchi, W., Flocchini, P., Casteigts, A., & Amblard, F. (2011). Time-varying graphs and social network analysis: Temporal indicators and metrics. *arXiv preprint arXiv:1102.0629*.

Shi, L., Wang, C., & Wen, Z. (2011). *Dynamic network visualization in 1.5D*. Paper presented at the Proceedings of the 2011 IEEE Pacific Visualization Symposium.

Shumway, J. M., & Otterstrom, S. M. (2001). Spatial patterns of migration and income change in the mountain West: The dominance of service-based, amenity-rich counties. *Professional Geographer, 53*(4), 492-502. doi:10.1111/0033-0124.00299

Slocum, T. A., McMaster, R. B., Kessler, F. C., & Howard, H. H. (2009). *Thematic cartography and geovisualization*: Pearson Prentice Hall Upper Saddle River, NJ.

Szego, J. (1987). Human cartography: Mapping the world of man. *Document-Swedish Council for Building Research*(D14).

Tobler, W. R. (1987). Experiments in migration mapping by computer. *American Cartographer, 14*, 155–163.

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science, 42*(sup1), 70-74. doi:10.1080/15230406.2015.1059251

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: can it facilitate? *International journal of human-computer studies, 57*(4), 247-262. doi:10.1006/ijhc.2002.1017

von Landesberger, T., Brodkorb, F., Roskosch, P., Andrienko, N., Andrienko, G., & Kerren, A. (2016). Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE transactions on visualization and computer graphics, 22*(1), 11-20. doi:10.1109/TVCG.2015.2468111

Wood, J., Dykes, J., & Slingsby, A. (2010). Visualisation of Origins, Destinations and Flows with OD Maps. *Cartographic Journal, 47*(2), 117-129. doi:10.1179/000870410x12658023467367

Yang, Y., Dwyer, T., Goodwin, S., & Marriott, K. (2017). Many-to-many geographically-embedded flow visualisation: An evaluation. *IEEE transactions on visualization and computer graphics, 23*(1), 411-420. doi:10.1109/TVCG.2016.2598885

Yang, Y., Dwyer, T., Jenny, B., Marriott, K., Cordeil, M., & Chen, H. (2019). Origin-Destination Flow Maps in Immersive Environments. *Ieee Transactions on Visualization and Computer Graphics, 25*(1), 693-703.

Zhao, J. F., Forer, P., & Harvey, A. S. (2008). Activities, ringmaps and geovisualization of large human movement fields. *Information Visualization, 7*(3-4), 198-209. doi:10.1057/palgrave.ivs.9500184

Zhu, X., Guo, D., Koylu, C., & Chen, C. (2019). Density-based multi-scale flow mapping and generalization. *Computers, Environment and Urban Systems, 77*, 101359. doi:10.1016/j.compenvurbsys.2019.101359

Appendix

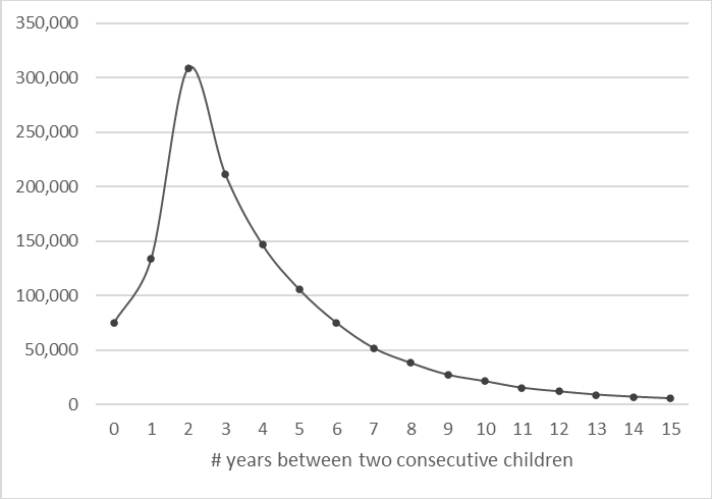


Figure A.1. The frequency of the number of years between two consecutive children.

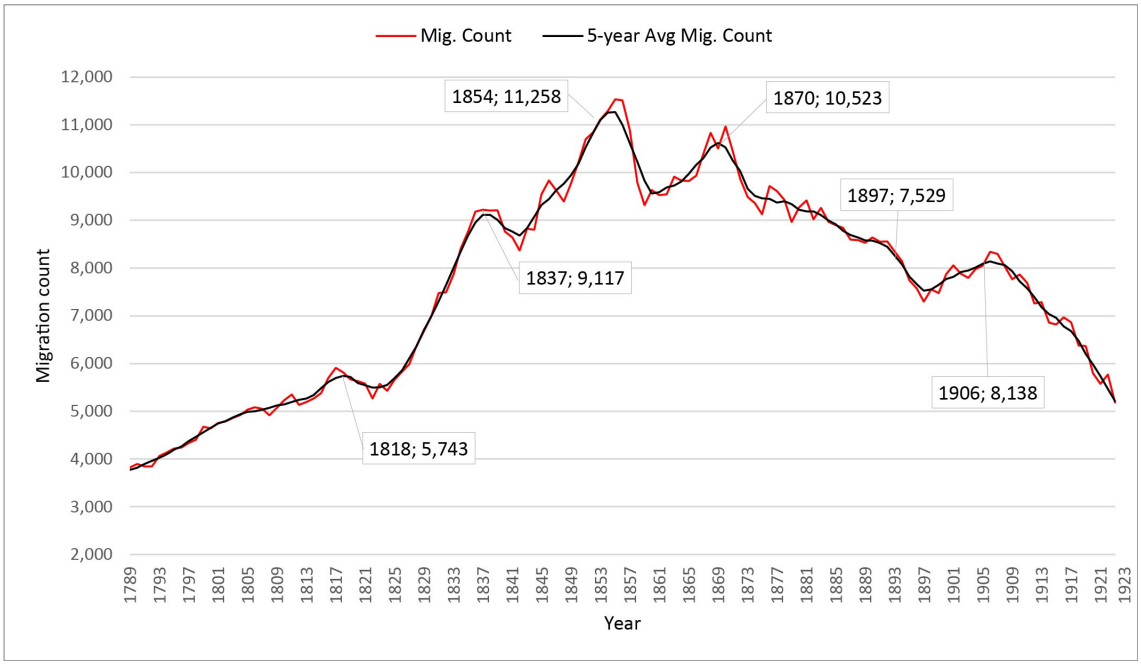


Figure A.2. Original family migration count (red) and smoothed migration count with 5-year moving average (black) between 1789 and 1924.

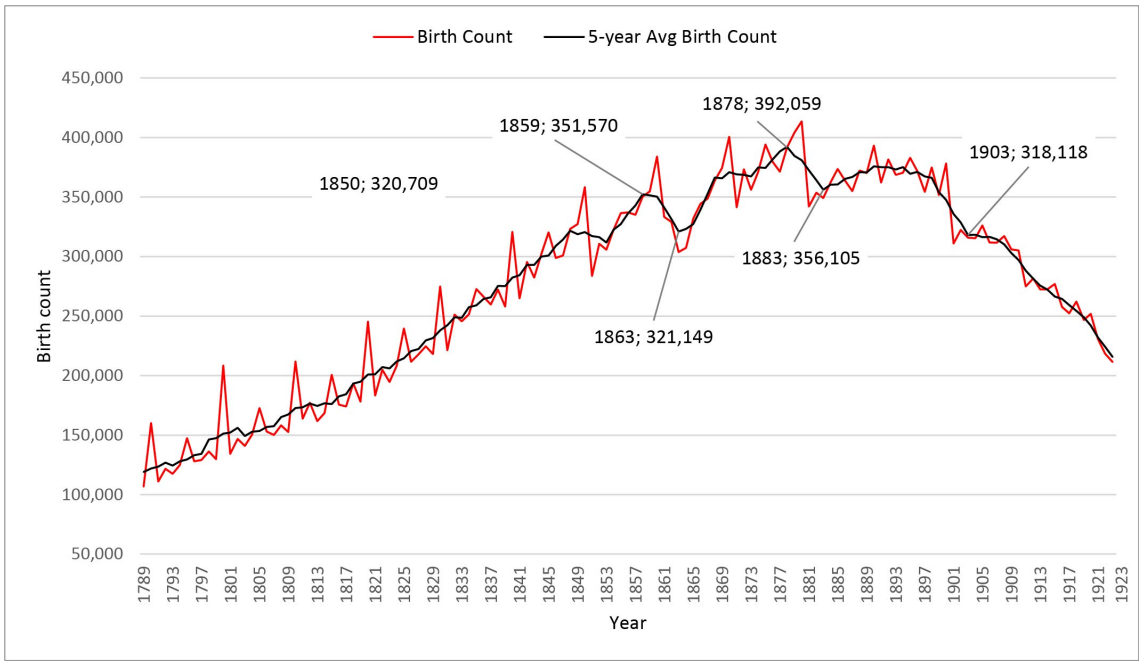


Figure A.3. Original number of births (red) and smoothed number of births with 5-year moving average (black) between 1789 and 1924.