

Supplementary Material
for
Large Scale Genealogies Distinguish Frontier from
Steady-State Internal Migration

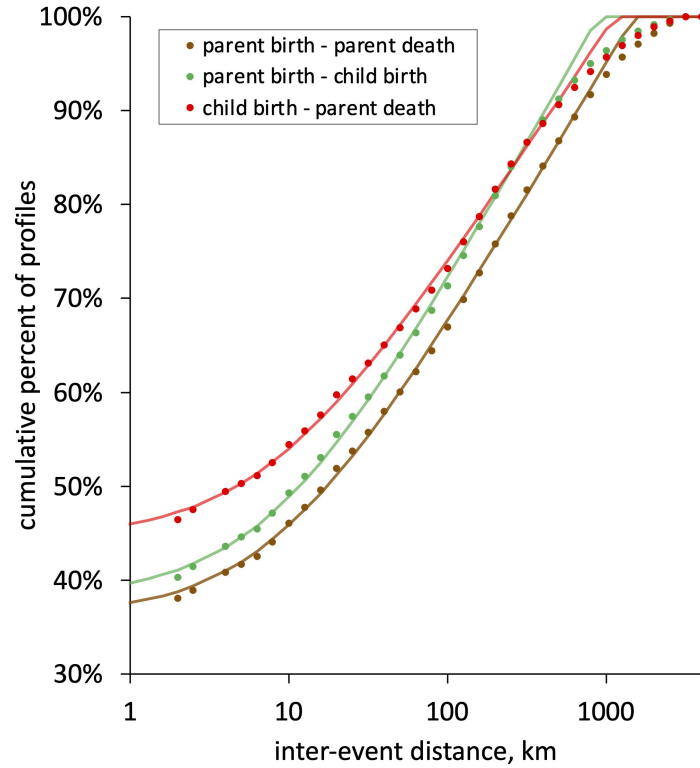
Robin W. Spencer and Samuel M. Otterstrom

In addition, animated maps and source code may be found at
<https://scaledinnovation.com/gg/migration/migration.html>

SM1: Validation of genealogic data

Prior Validation Our principal dataset has been extensively curated and validated by its authors (Kaplanis et al. 2018) and examined by others (Chong et al. 2022, Stelter and Alburez-Gutierrez 2022, Colasurdo & Omenti 2024). Principal criticisms involve small differences in mortality statistics in Scandinavia; our concerns are with location and distance, not age at death, and we principally examine North America with far higher sampling.

Birth-Death Distance as Migration Distance At an individual level, migration distance is somewhat ambiguous: is it the single largest distance that someone may move in a lifetime? Is it the summed magnitude of their moves, or the vector sum? We have used genealogical birth and death locations to measure migration, which are appropriate as the literal beginning and end of a lifetime, as well as available in genealogies. Since the Kaplanis dataset includes (anonymized) parent and child identifiers, we can also examine waypoints, namely the births of any children. The distributions below, for distances less than 4000 km (i.e. not trans-continental) are for parent birth to parent death distance ($n = 828,000$), parent birth to first-child birth ($n = 854,000$) and first-child birth to parent death ($n = 893,000$). The only significant difference is that a smaller percent move from first-child birth to parent death ($f_{leave} = 0.55$) than parent birth-death ($f_{leave} = 0.64$) or parent birth to first-child birth ($f_{leave} = 0.62$), as expected if most migration occurs before young adults have children. In addition, the chart shows the typical quality of fit for the cumulative genealogic data (dots) to Equation 1 (solid lines).



The trends observed of subsets of birth-death distances, such as the r_{fixed} spike in the US 1750-1890 or time-trends in r_{max} (Figure 4) also appear in the parent-child distributions (not shown).

Regional Bias Any regional or selection bias — for example, if present-day Virginians disproportionately posted online family trees compared to Vermonters — is reduced by averaging and fitting in smaller geographic blocks, then comparing and mapping across blocks. Our conclusions are based on block-averages, normalized percents, and model fitting, not on absolute numbers.

Time Bias The dataset contains more profiles of people born 1920-1950 than other generational slices, since online genealogy customers know their parents and grandparents better than earlier ancestors. But as with geography, such bias is decreased by averaging and fitting to smaller blocks of time, and drawing conclusions from ratios and fits rather than absolute

numbers. All subsets used for mapping, averaging, or curve fitting contain at least 100 ancestor profiles. We generally time-slice by thirty years to match generation times, which are also readily measured in the Kaplanis dataset: women = 29.7 ± 7.1 years, $n = 762,000$, and men = 34.4 ± 8.8 years, $n = 916,000$.

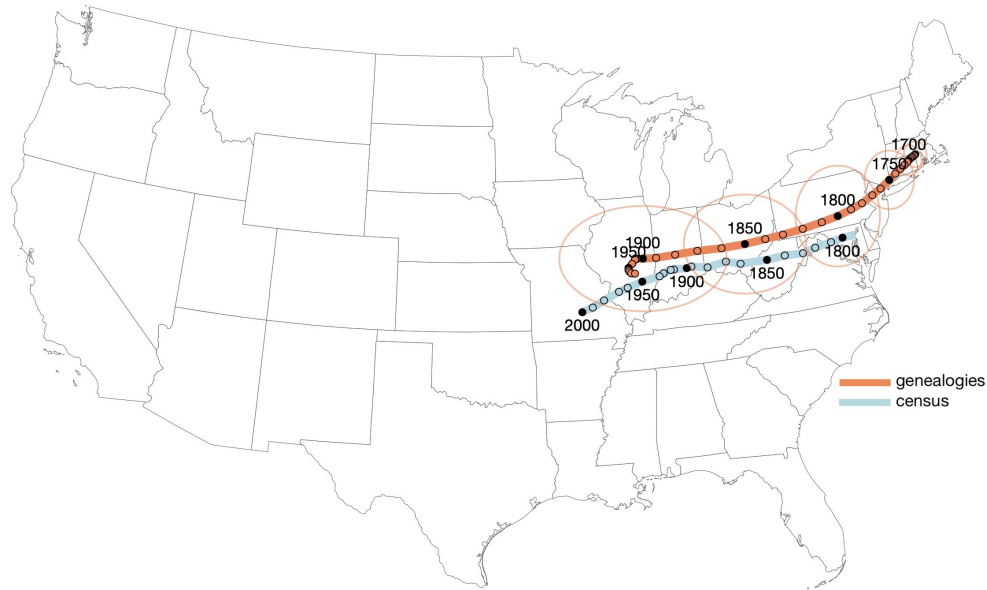
Gender Bias Genealogic information will inevitably have a gender bias in a patrilineal culture: surnames do not provide continuity for female branches and ancestors, making them more difficult to trace. Our gender-based analyses are therefore not across but within genders, e.g. SM2, SM3 below.

Generic Location Error Kaplanis et al report birth and death locations as latitude and longitude which they compute by automated geocoding of user-entered place names. In our context, a nation-scale descriptor, for example “Britain” or “Canada” is unacceptably vague, while a town or county level descriptor is useful. We identify nation-scale profiles as very high-density blocks in central unpopulated areas, and ignore them in our tallies.

Specific Anomalies Any crowdsourced genealogy will contain anomalies; algorithms can identify only some types of error, i.e. children older than parents. In Kaplanis we find, for example, reported ancestry in Oregon in the early American colonial era, which is either simply incorrect or may indicate an estimate of Native American ancestry. Rather than impose judgment, we generally require all sampled time-and-space data subsets to contain at least 100 profiles and use robust measures, median or geometric mean, that are less sensitive to outliers especially with a power-law distributed measure like migration distance.

Comparison to Census Results The center of population, computed from census data, is an iconic measurement in demographics (https://www2.census.gov/geo/pdfs/reference/cenpop2020/COP2020_documentation.pdf). It may also be estimated from genealogies: we count everyone alive in each tally year (each decade, to match US census data), and note their location as their birth location if under age 20 and death location if older. The result, shown below, is a time-track for the US center of population that extends back to the early colonial era, well

before the first census in 1790. The orange ellipses extend one standard deviation in latitude and longitude from each center point and show that the genealogic and census population centers coincide, well within one sigma, from 1800 to 1950, after which, as noted elsewhere, genealogic data become very sparse.

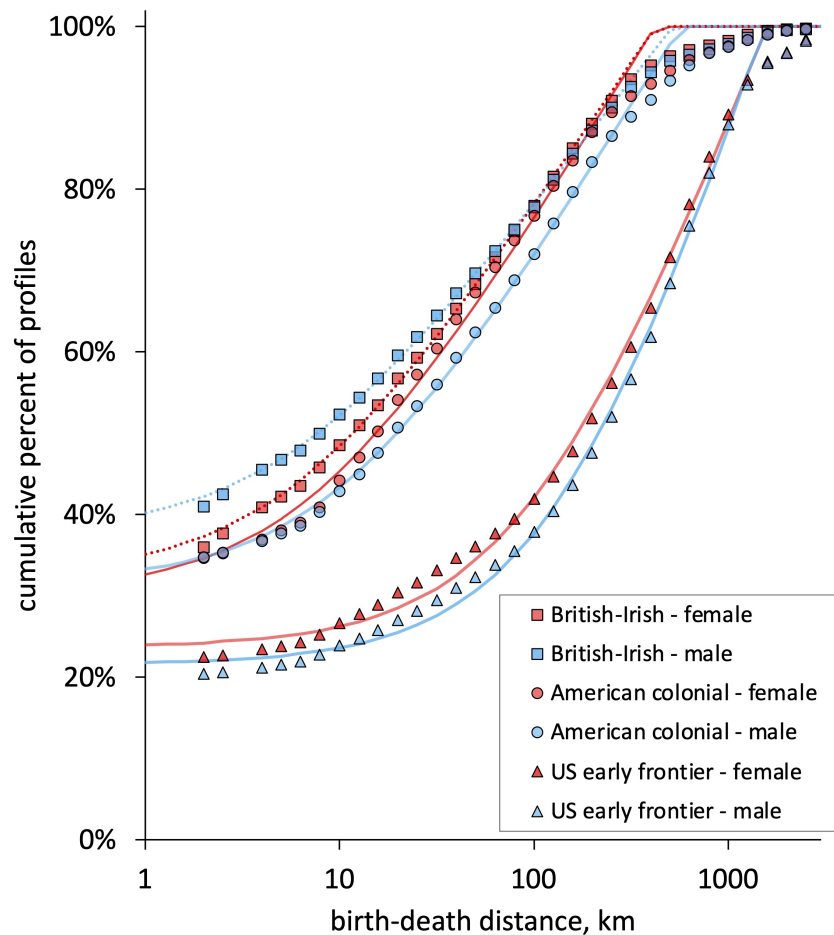


This is a stringent comparison since it requires an estimate of total population (rather than percent or model fit) and is therefore subject to any regional bias. The systematic 0.7° latitude northerly difference between genealogic and census centers (1800-1900) probably reflects such a bias, which is consistent with the much greater density of historic societies north of this line (Rosenstein & Vakharia 2022, Figure 4) which support a deeper record-keeping of genealogic information.

Many other maps are available; in particular the movement of the frontier, as measured by population density (e.g. <https://www.census.gov/dataviz/visualizations/001/>), is also illustrated by the transition from migration to settlement <https://scaledinnovation.com/gg/migration/migration.html?type=origin&loop>.

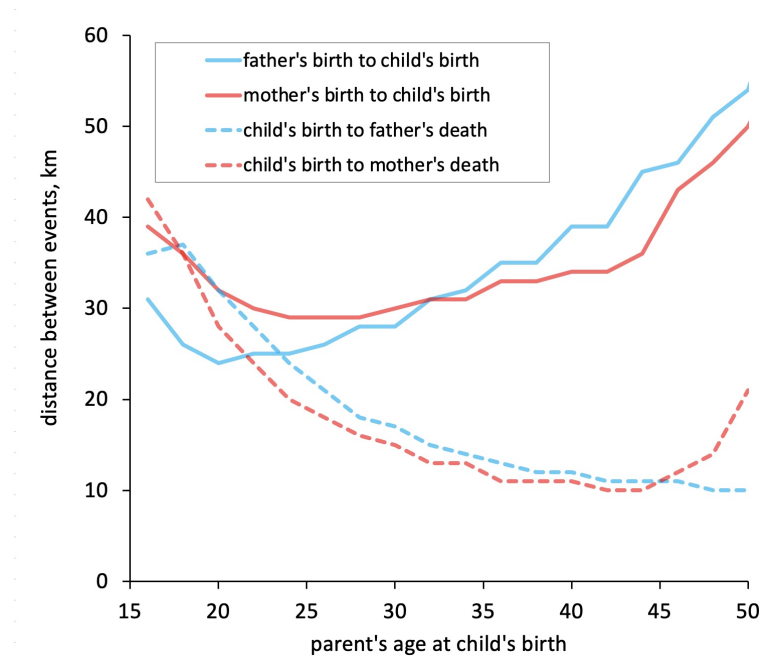
SM2: Dependence on Gender

Any analysis of genealogic data must be concerned by gender bias, since in a patronymic society people are more successful in identifying male ancestors because of the persistence of their surnames. The Kaplanis profiles identify gender so we can analyze female and male subsets separately and compare the results. The chart below shows distributions by gender for 1.5 million birth-death distances in three subsets (Britain and Ireland (all dates), female $n = 174,143$, male $239,192$; American 1620-1770 female $216,023$, male $269,400$; American 1770-1860 female $255,475$, male $354,906$). Least-squares fits to Equation 1 are shown as solid (American) or dashed (British-Irish) lines. Any gender differences are small compared to the differences that interest us, namely those across time or geography.



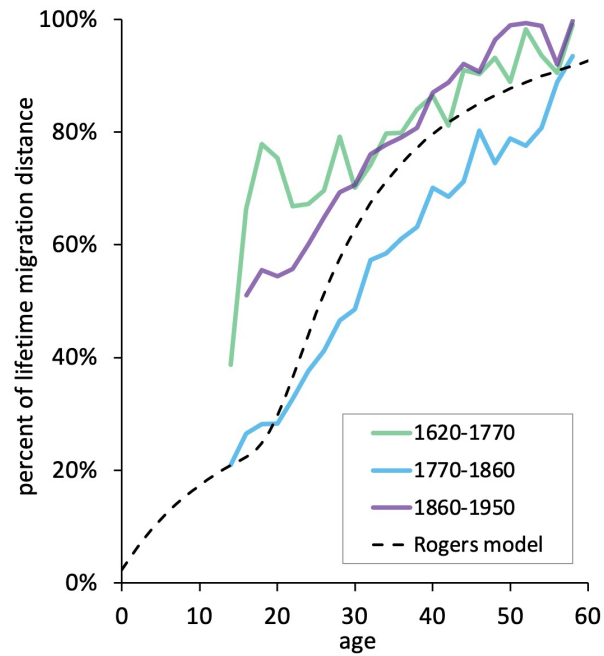
SM3: Age at migration

The age at which people move is relevant to a discussion of their migration, but is not directly available from genealogic data that record only birth and death. But since the Kaplanis data include parent-child links (via anonymous id numbers), we have dates and locations for intermediate life-events, namely the births of children, which we analyze in three ways. A child born before their parent's migration will be closer to that parent's birthplace than death-place, and vice-versa. The chart below shows geometric means of the distance between genealogical events for 1.66 million profiles, all regions and dates. The crossover point at which the (parent birth)-(child birth) distance equals the (child birth)-(parent death) distance should occur at the parent's age of migration; this occurs at about age 18 for women and age 23 for men, consistent with migration in early adulthood.



Births of multiple children are milestones along a parent's life-path. The results below are from 27,800 profiles of American internal migrants, all with four or more children constrained to a near-linear path from parent birth, to all child-births, to parent death. These data are generally consistent with Rogers' consensus age-of-migration model (Rogers & Castro 1981, table 17, shown below as its cumulative distribution) which peaks at age 23, though

the peak age in early frontier era (1770-1860) is about 10 years later than in the colonial era (1620-1770) or later period (1860-1950).



Thirdly, as noted above (SM1, Birth-Death Distance as Migration Distance), the distance distributions between a parent's birth, death, and their first child's birth, are also consistent with migration in young adulthood.

SM4: Three-parameter model

Based on a Zipf-like view that migration cost is proportional to distance and the simplicity of the distributions of Figure 1, the basis for any mathematical description is

$$p_{migrate} \approx 1/r$$

“Cost” could also include a linearly adjustable distance-independent component

$$cost \approx (1 - f) + f \cdot r; f = 0 \dots 1$$

When $f \rightarrow 0$, $cost \rightarrow 1$, and when $f \rightarrow 1$, $cost \rightarrow r$. Adding an independent stay-leave parameter (because only the leavers pay a moving cost) and replacing r with this cost function gives

$$p_{migrate} \approx (1 - f_{leave}) + f_{leave}/(1 - f + f \cdot r)$$

Integrating and substituting $r_{fixed} = 1/f$ to make the logarithm scale-free and cast the fixed-cost parameter into distance units, gives

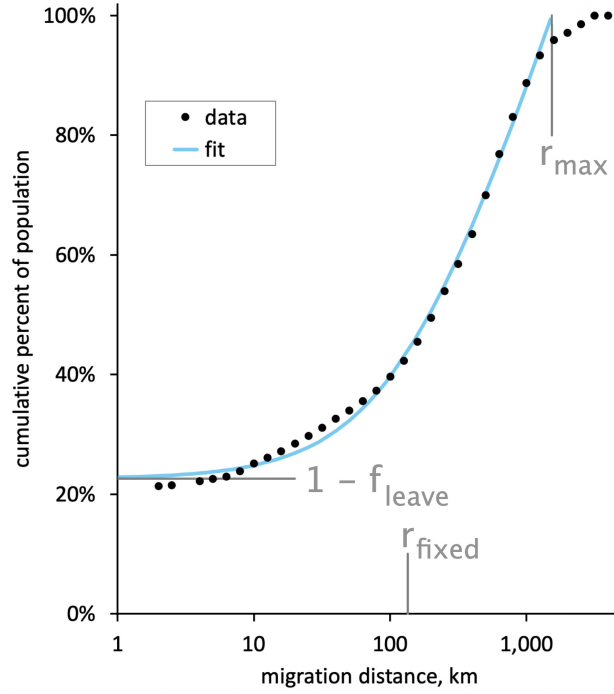
$$CDF \approx f_{leave} \ln(1 - 1/r_{fixed} + r/r_{fixed})$$

Removing the scale dependence inside the logarithm and making the assumption that interesting distances are $\gg 1$ km gives

$$CDF \approx f_{leave} \ln(1 + r/r_{fixed})$$

Normalizing so that $CDF \rightarrow f_{stay} = 1 - f_{leave}$ as $r \rightarrow 0$ and $CDF \rightarrow 1$ as $r \rightarrow r_{max}$ gives

$$CDF \approx 1 - f_{leave}(1 - \ln(1 + r/r_{fixed})/\ln(1 + r_{max}/r_{fixed})) \quad (1)$$

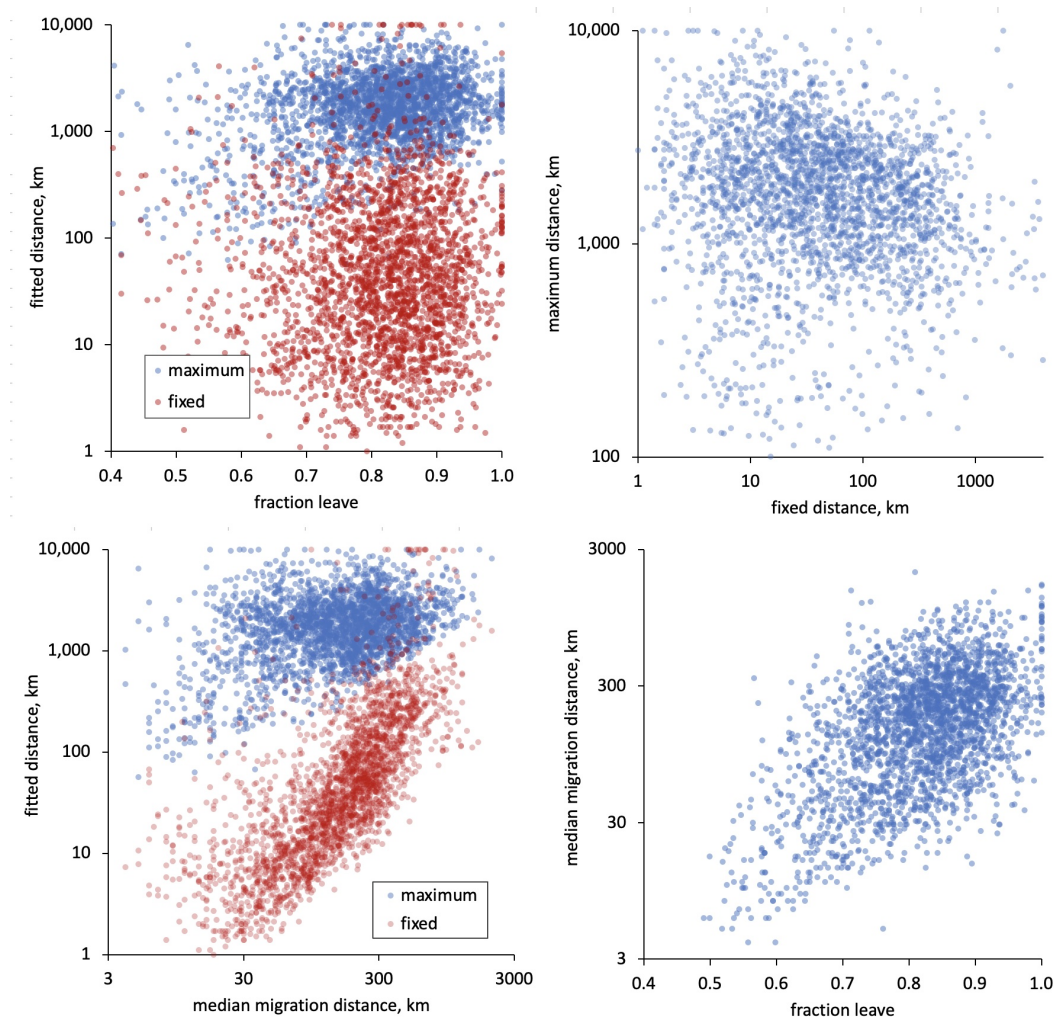


The interpretation of these parameters is shown above with the least-squares fit to a typical dataset (US 1770-1860; 608,000 profiles), in this case with fitted parameters $f_{leave} = 0.774 \pm 0.002$, $r_{max} = 1532 \pm 27$ km, $r_{fixed} = 135 \pm 5$ km, rms error 1.3%. Other algebraic forms (additive, multiplicative, reciprocal, etc.) were tried for a cost parameter with poorer fits and greater systematic differences from the observed data. Equation 1 is justified by its simplicity, facile interpretations for f_{leave} and r_{max} , and fits to observed data. It is not derived by analogy to inanimate physics.

SM5: Model Parameter Independence

Even though our model has only three fitted parameters, it is important to see if they are correlated. We created a test dataset by partitioning North America into one degree latitude by longitude blocks, then tallying all profiles for each block for those age 20 within each 30-year time slice from 1620 to 1950. Only subsets with at least 100 profiles were kept ($n = 2577$) and their birth-death distances fitted to Equation 1.

The parameters are uncorrelated (graphs below: top row). Median birth-death distance is a robust observable (not a fitted result) and is correlated with all three fitted parameters : r_{fixed} (strongly), r_{max} and f_{leave} (less so) (bottom row).



SM6: Standardizing Migration Data

The diversity of sizes of administrative areas has long frustrated attempts to compare percents migration across countries and epochs (Long & Boertlein 1976). When the distance between administrative areas is large, the underlying $1/r$ distribution of migration distances results in an underestimate of those moving. Viewing migration distances as a distribution and fitting that distribution to a model, e.g. Equation 1, can reduce the bias in the simplest migration question: How many people moved?

Two datasets can be compared if they have the same resolution limit, which is the smallest accurately measurable distance. For the Kaplanis genealogy data, that limit is, empirically, 10-20 km = r_{fixed} for places and time periods except the clear American anomaly 1760-1890 (Figures 4, 5). This is consistent with a simple estimate of the average distance between towns: contiguous US area $\approx 8,000,000 \text{ km}^2$, 20,000 incorporated villages, towns, and cities gives 40 km^2 per town, thus $\sqrt{40} \approx 6 \text{ km}$ between neighboring town centers. In other words, the observed resolution coincides with the expectation that genealogic event recording is at the scale of towns and villages. Most migration studies depend on data collected by administrative area — county, state, or country — in which case the resolution limit is set by the distance between those areas. Given the $1/r$ distribution, in which there are many more short jumps than long jumps, an average distance between the centers of nearest-neighbor administrative areas is a good measure of the resolution limit.

Datasets with different resolution limits can be compared either by data truncation or by distribution extrapolation. In truncation, migration distances for the finer-grained dataset below the longer limit are set to zero; i.e. if the coarser limit is 100 km, a person who moves 75 km is judged not to have moved at all. Extrapolation involves extending the $1/r$ trend (= log trend in the CDF) of the coarse dataset down to the limit of the fine-grained dataset. Both methods agree for a comparison of the US 2010 census data (tallied by commuting zones with a resolution limit $\sim 80 \text{ km}$) to the Kaplanis data, and show that young-adult tendency to leave home in 2010 was comparable to what it was in 1710, which is to say 50-60% instead of the uncorrected 30% (SM7).

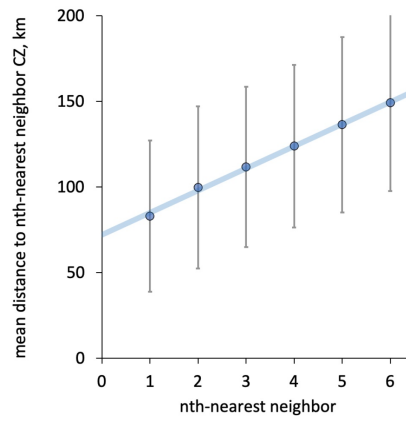
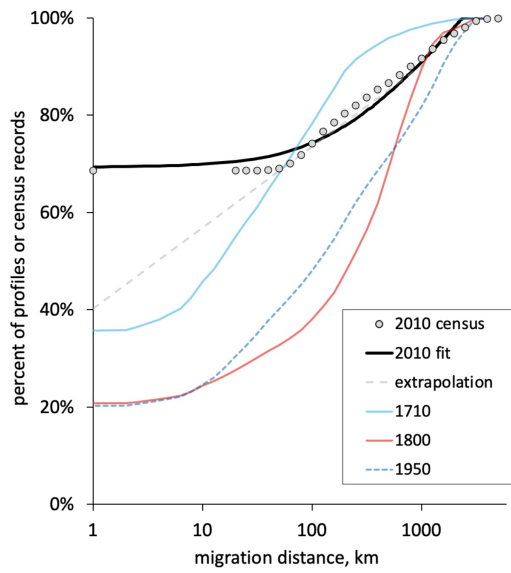
SM7: 2010 US Census Migration Data

Since the genealogy-derived Kaplanis data do not extend much past 1950, any data that illuminate recent behavior is of interest. Recent US data are available for 31 million young adults in 2010 (Sprung-Keyser et al. 2022), and as shown in Figure 2a, also display the $1/r$ distribution of migration distance. The CDF is shown below-left, overlaid for comparison with the Kaplanis data from 1710, 1800, and 1950 which are also shown in Figure 2.

The 2010 data measure the locations of people born between 1984 and 1992 as tallied in 2000 and 2010 census and taxation data, essentially measuring where they were at age 16 and then at age 26. This is not the birth-to-death distance of the Kaplanis data but should capture the peak years for migration.

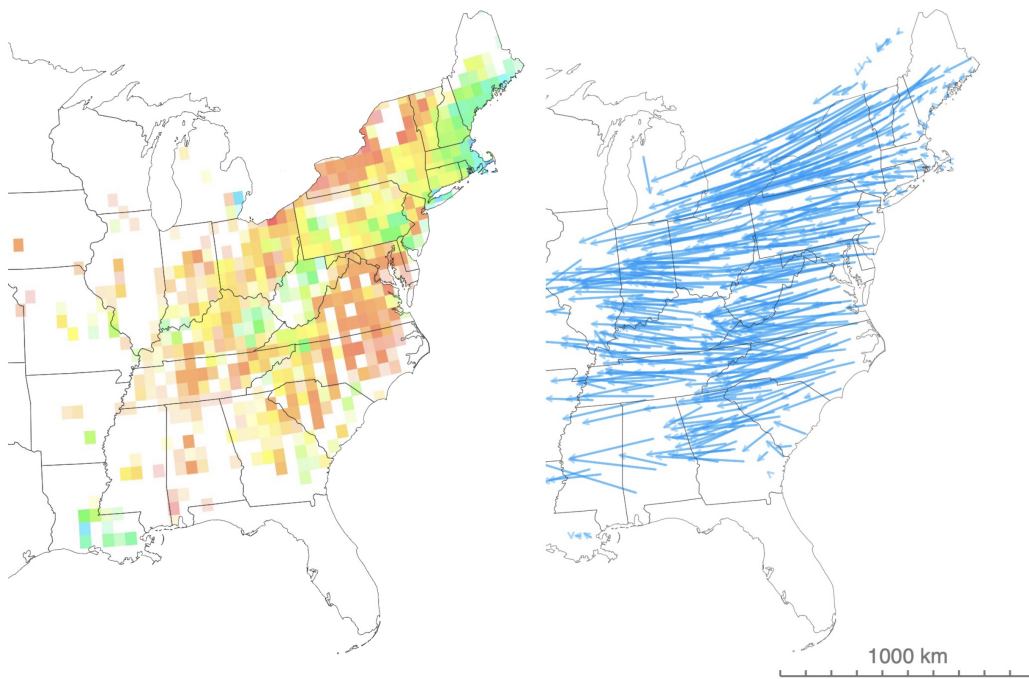
The 2010 data have a fitted value of r_{max} of 2380 km, at the plateau since 1830 (Fig 4b) and which simply reflects the maximum possible migration distance (the numbers moving between the contiguous states and Alaska or Hawaii are small). The vertical intercept, $f_{stay} = 1 - f_{leave}$, is 70%, well above the 20-35% range of the Kaplanis datasets, but we do not think that this implies that people stay close to home much more in recent years than in earlier decades. Instead, this reflects the distance resolution problem when comparing different data sources. Kaplanis locations are latitude and longitude from geolocation, and therefore largely at town-level resolution. Sprung-Keyser is coarse, placing a person in one of 741 commuting zones (CZ); as shown below-right, the mean distance from the geographic center of a CZ to its nearest neighbor is 83 ± 44 km, exactly at the sharp break in the CDF (gray dots, below left). Simulated truncation of continuous data, for example setting all distances below 80 km to zero, gives the same sharp transition from a $1/r$ distribution to a flat plateau.

Extrapolating the $1/r$ region from ca. 80 km to 1 km resolution (gray dashed line, below left) suggests that a value of f_{stay} of about 40% is more comparable to the Kaplanis datasets; in other words modern stay-vs-leave behavior has reverted a range typical of pre-industrial Europe or pre-frontier America. Truncation gives the same result, since the census data (gray dots) overlaps the colonial distribution (light blue) at the former's resolution limit of 80 km.

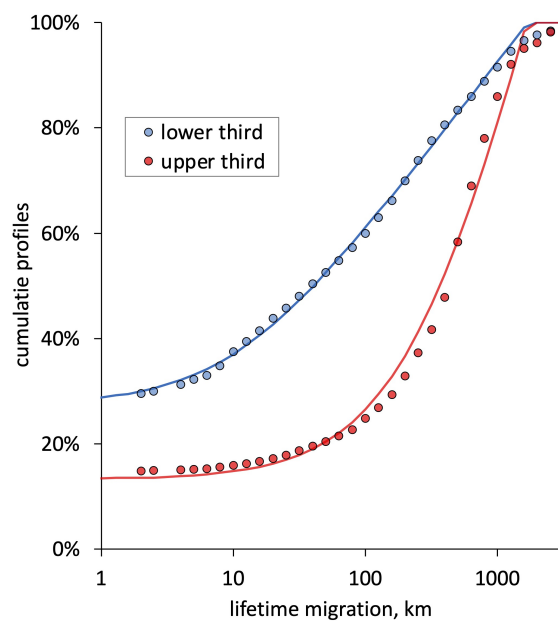


SM8: Geographic Dependence of the Fixed Distance Parameter

The r_{fixed} term in Equation 1 displays a unique spike in the American data from 1770-1860 (Figure 5), with values of 50-250 km that cannot be explained as a resolution limit in the dataset. The map below-left shows the median values of fitted r_{fixed} in this era for every $0.5^\circ \times 0.5^\circ$ block with at least 100 profiles. A geographically distinct pattern is apparent with low values on the New England coast and very high values in western New York and inland Virginia to South Carolina. The lower chart shows the data (dots) and fits (lines) for the lower and upper third of these geographic blocks; the distribution for the low r_{fixed} blocks (blue) is very similar to that for all colonies 1620-1710 (Figure 3) with a higher r_{max} .



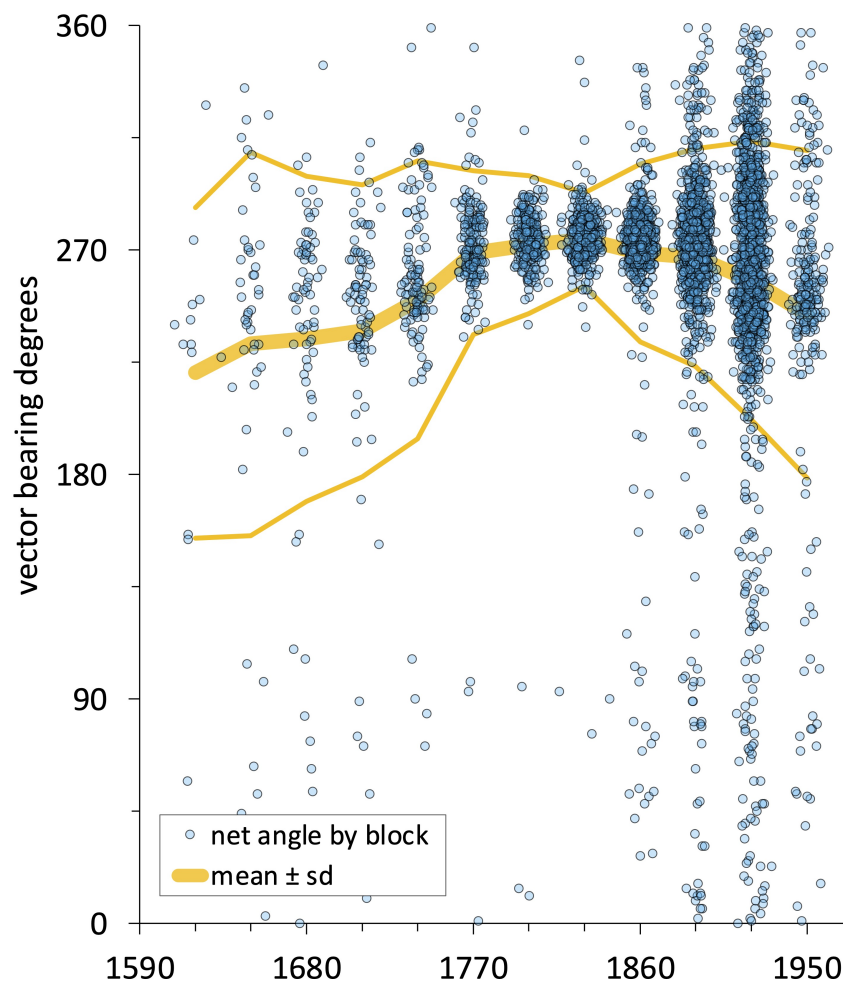
A physical explanation is suggested by the vector sums of all migrations for the blocks (above-right): most migrations from the high- r_{fixed} blocks have infeasible or undesirable barriers to the west, either Ontario or a Great Lake, or the Appalachian Mountains. With reports that better land was available across those barriers, the decision to migrate may have had an implicit minimum fixed distance.



We note that in a closed system, the combination of high r_{fixed} and f_{leave} would be self-extinguishing: most children would move a considerable distance away from home. American internal migration away from the east was, however, back-fueled by a very high net survival rate as well as high transatlantic immigration.

SM9: Directionality of American Migration

The directionality of American internal migration 1620-1950 is shown below. Each blue dot is the bearing (0 = North, 90 = East, etc) of the vector sum of all birth-death vectors for all 0.5° square geographic blocks in North America in the given 30-year slice, with a minimum of 100 profiles. Dots are horizontally dithered to show their distribution; orange lines are the mean $\pm$ standard deviation of the bearing for each time period.

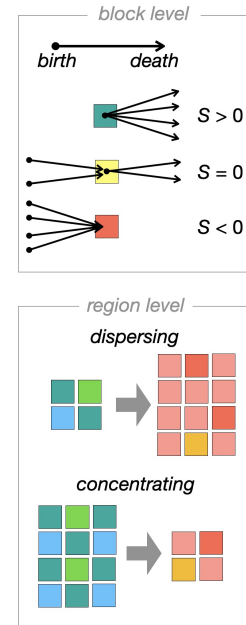


SM10: Migration Symmetry

Migration symmetry S is computed for a geographic block (typically a county or latitude-longitude block) as

$$S = (\text{births} - \text{deaths}) / (\text{births} + \text{deaths})$$

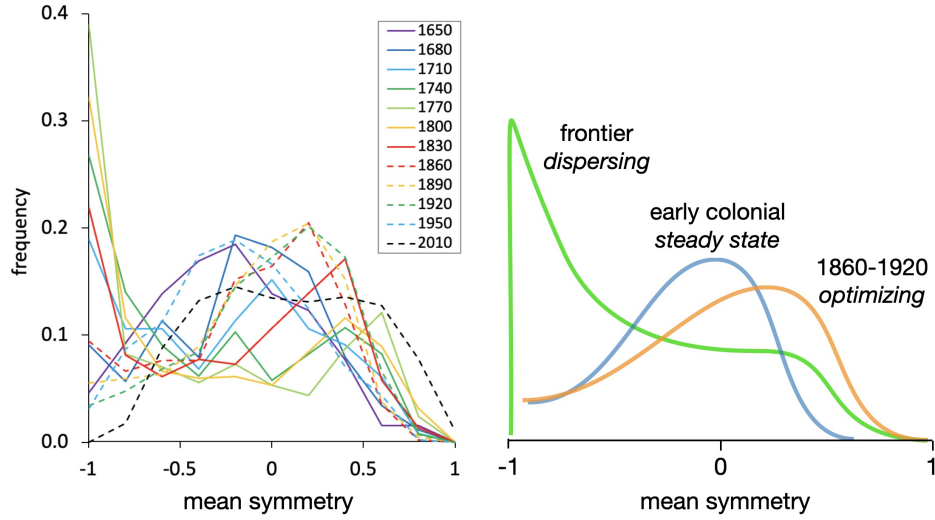
where births = number of genealogic profiles with birth location in the given block but death outside, and deaths = number born outside the given block but died within. Symmetry is a migration metric, thus within-block births and deaths are ignored. If S equals zero, people may be moving but influx equals efflux and the block is at a migration steady-state. The red-to-yellow-to-blue scale at the right is the same used in the maps below.



At larger scale, if a few blocks lose people to a larger number of blocks ($\sum S < 0$), the population is dispersed and diluted. This is typical of an early frontier in which the destination area is larger than the source area. Urbanization is the opposite, in which a larger number of usually geographically disperse blocks feed a smaller number and the population becomes concentrated ($\sum S > 0$).

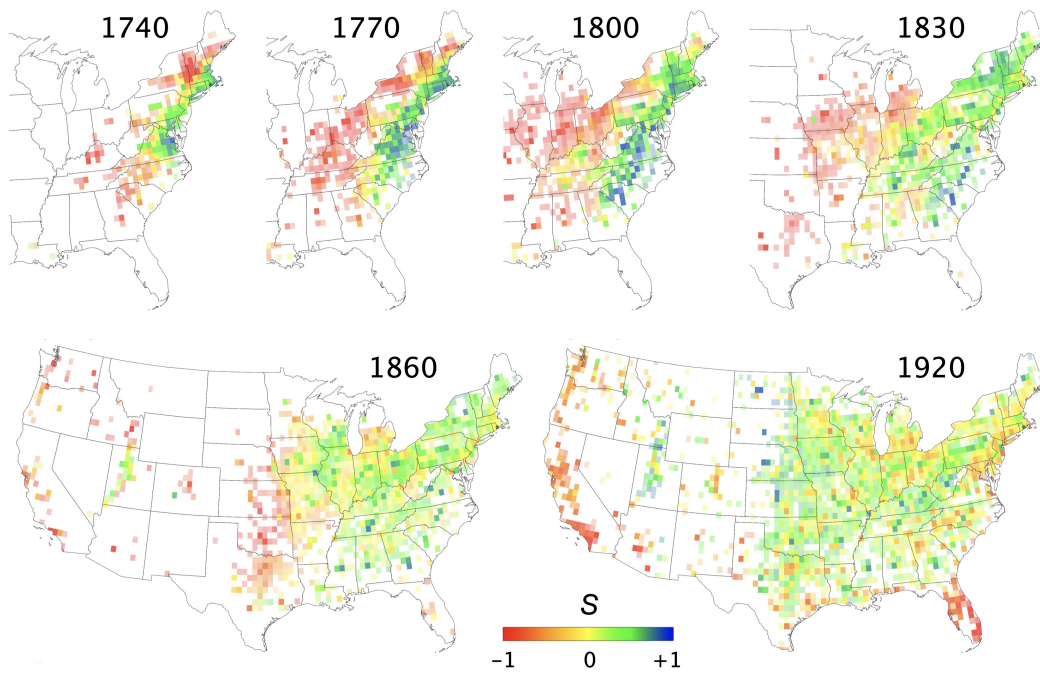
Just as for the distance metrics, American migration is well described by both distributions and maps of symmetry over time.

The distributions (below; left: data, right: illustrative sketch) are of the means of S for all $0.5^\circ \times 0.5^\circ$ blocks with $\text{births} + \text{deaths} > 100$ (number of blocks $n=25$ for 1620, $n=1717$ for 1920). Until ~ 1680 , the colonies were close to a steady state with a centrosymmetric distribution of S . The 1710 map (see SM11) shows the first development of a frontier, apparent in the distributions as a sharp rise in blocks with $S \approx -1$, i.e. unidirectional migration into new territory, a trend which continues for 5-6 generations.



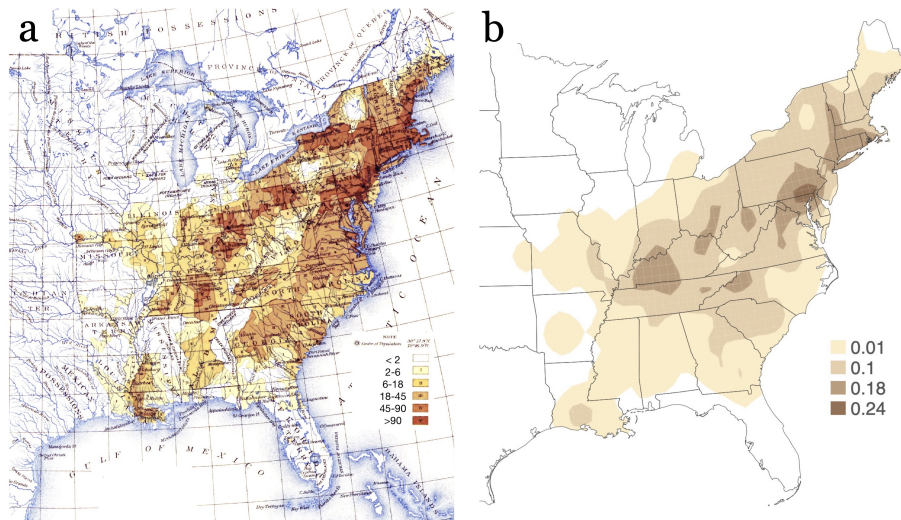
By 1830 the distribution becomes bimodal with a preview of western urbanization (Chicago, Salt Lake City, San Francisco, and Los Angeles). For 1860-1950 a steady-state dominates east of the Mississippi (yellow and light green) and the pronounced asymmetries are urbanization and moves to California and Florida. Census data for 2010 commuting zones (Sprung-Keyser et al 2022) show a return to a centered distribution but with a mean absolute symmetry of 0.38, still very dynamic.

While non-zero symmetry S highlights different situations, frontier migration is distinguished by being highly directional, as shown by the strong east-west color gradient 1710-1830. Urbanization pulls from diverse locations and is not uniquely directional.



SM11: Alternative Frontier Mappings

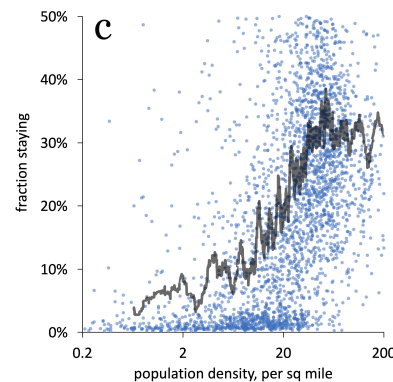
The American frontier is classically depicted as contours of population density based on decennial censuses since 1790, such as (a) Gannett's map for 1830 (Gannett 1903) and its online equivalents at <https://www.census.gov/dataviz/visualizations/001/>. We offer two alternatives from genealogic data.



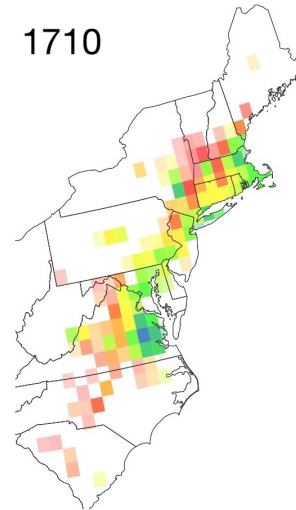
A reasonable match (b) is seen with contours of $f_{stay} = 1 - f_{leave}$, obtained by fitting Equation 1 to all profiles in each $0.5 \times 0.5^\circ$ block.

Chart (c) shows the relation between population density and f_{stay} for 3200 counties' census data 1790-1890 and 100 or more genealogy profiles per county in each decade.

The relation is monotonic and shows a rapid rise in “staying” behavior at 6-20 people per square mile, albeit with considerable scatter (the dark line is a smoothed running average). The traditional Gannett frontier line corresponds to this rise, i.e. the first few percent of those who don’t move. Like population density, f_{stay} is a threshold metric: it does not revert but generally keeps rising as the frontier passes by.



Birth-death symmetry S is a more information-rich descriptor for the frontier line. In our maps, profiles are assigned to locations at age 20 (SM1, SM5), thus areas with negative mean symmetry (orange and red in the maps) are a view to the future, where these people died ~ 20 -50 years after the map date. The map for 1710 (right) is nearly three generations before the first 1790 census data. We see the frontier line in yellow (i.e. $S \approx 0$) as well as the future of those crossing that line 1710-1740, namely settlement into Vermont, New Hampshire, eastern New York, and movement west from the Tidewater region.



The symmetry map for 1830, below, bears comparison to the Gannett and f_{stay} maps above. Again the $S \approx 0$ boundary corresponds to the traditional frontier line, consistent with the expected transition between dispersive and concentrative migration. The future-view clearly shows these migrants' destinations, including the beginnings of western urbanization in Chicago, Salt Lake City, San Francisco, and Los Angeles.

Complete 1620-1950 maps for these and other metrics are at <https://scaledinnovation.com/gg/migration/migration.html>.

