THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE

Economic History Working Papers No: 385

Crowd-sourced Chinese genealogies as a tool for historical demography

Melanie Meng Xue, LSE

October 2025

Historical Economic
Demography Group
Research at LSE

Economic History Department, London School of Economics and Political Science,
Houghton Street, London, WC2A 2AE, London, UK. T: +44 (0) 20 7955 7084.

Crowd-Sourced Chinese Genealogies as a Tool for Historical Demography[*]

Melanie Meng Xue[†]

LSE

September 2025

Abstract

This paper introduces a structured approach for using genealogical records from FamilySearch to study Chinese historical demography. As a proof of concept, we focus on over 190,000 digitized records from a single surname, drawn from many provinces and spanning multiple centuries. These lineage-based microdata include individual-level birth, death, and kinship information, which we clean, validate, and geocode using consistent rules and standardized place names. We begin by documenting descriptive patterns in population growth, sex ratios, and migration. Migration was overwhelmingly local, with long-distance moves rare and concentrated in a small number of lineages. Out-migration rose to a high point between 1750 and 1850 and then declined in later cohorts and generations. We then use the genealogical data to test specific hypotheses. Male-biased sex ratios—likely influenced by female infanticide—are strongly associated with higher rates of male childlessness. Migration rates fall sharply with patrilineal generational depth, offering micro-level evidence that clans became more sedentary over time. Together, these findings show how genealogical records can be used to reconstruct long-run demographic patterns and to assess social processes such as kinship, mobility, and reproductive exclusion. The approach is replicable and extensible to other surnames and regions as data coverage improves.

Keywords: Crowd-sourced genealogies, Historical demography, China
JEL Codes: J11, J13, N1, N35

[*] I am grateful to Wolfgang Keller and Carol Shiue for their helpful comments on this project. I also thank a colleague for thoughtful suggestions that clarified the argument and improved the exposition. I especially thank Yan Li and Yuyang Zhang for excellent research assistance, and I acknowledge Joey Li and Youyang Zhang for their contributions as research assistants at an earlier stage of the project, when a different approach was being investigated.

[†] Assistant Professor at London School of Economics and Political Science (LSE) and Research Affiliate at the Centre for Economic Policy Research (CEPR). Email: m.m.xue@lse.ac.uk

I Introduction

Chinese lineage books (*zupu*) are among the longest and most detailed continuous population records in the world. They list births, deaths, marriages, and places of residence for dozens of generations, offering a window on demographic behavior long before modern censuses. Until recently, however, most surviving genealogies were locked away in temple chests or scattered local archives, limiting researchers to small, hand-collected samples. FamilySearch has begun to change this landscape: the non-profit has digitized millions of Chinese lineage pages and posted the transcribed entries—complete with links between parents, children, and spouses—on an open-access web platform. The sheer scale of these data creates new opportunities for historical demography, but it also raises practical hurdles: duplicated entries, inconsistent place names, and highly uneven geographic coverage, any of which can bias naïve analyses.

A parallel line of work in economics and economic history uses *crowd-sourced genealogies*—large, collaboratively maintained family trees—to recover long-run demographic and social patterns. Platforms such as Geni, Ancestry, MyHeritage, and the FamilySearch Family Tree enable record linkage across generations at scale, complementing or substituting for censuses where unique identifiers are missing. At global scale, researchers have mined Geni's multi-million-person tree to study marriage distances, migration, assortative mating, and longevity over five centuries (Kaplanis et al., 2018). In the United States, genealogy-assisted linkages ("Census Tree") combine the Family Tree's user-contributed kin links with machine learning to connect historical census records, yielding new panels for work on intergenerational outcomes (Price et al., 2021; Buckles, Haws, et al., 2023; Buckles, Price, et al., 2023). Demographers have also evaluated data quality and selection in aggregated online trees such as FamiLinx (Colasurdo and Omenti, 2024). We build on this literature by assessing how far *crowd-sourced* Chinese genealogies can recover core indicators once cleaned and validated.

This paper sets out a workflow for converting raw genealogical records from FamilySearch into a research-ready panel of individuals and lineages. We illustrate the procedure with a single—but very large—surname sample: 李 (Li). As one of the most common surnames in China, Li accounts for about 7.2 percent of the population (Ministry of Public Security, 2019). We (i) scrape and de-duplicate individual records bearing the character "李" or its Latin transcription, (ii) stitch them into extended family trees anchored on a common founding ancestor, (iii) standardize place names using the China Historical GIS, and (iv) apply a set of validation rules to flag likely errors. The resulting dataset contains 192,310 unique individuals spanning fifteen centuries and every province of China.[1]

Although the evidence comes from only one patriline, it is rich enough to replicate classic findings and to suggest new ones. We confirm the well-known southern bias in male-to-female ratios; we document overwhelmingly local mobility with a thin, heterogeneous long-distance tail; and we show that out-migration falls sharply after the first few generations of a lineage, consistent with growing attachment to ancestral places. Most importantly, the exercise demonstrates that large genealogical datasets can be studied with the same statistical care routinely applied to census or survey microdata. The tools we describe—transparent scraping scripts, geographic reconciliation, and distance-based migration coding—are released with the paper and can be applied surname by surname until the broader universe of Chinese lineages is mapped.

By offering both a methodological framework and substantive results, the study aims to lower the entry cost for economists, demographers, and historians interested in China's long-run population dynamics, and to highlight the analytical promise of a source that has so far been used mainly for local case studies.

II Genealogical Sources in Chinese Economic History

Chinese genealogies (族谱, 宗谱) have become a valuable source for historical demography and economic history, particularly as digitization and improved methods have expanded access to these records. While their use began with anthropological and qualitative analyses of lineage behavior (Freedman, 1958, 1966), the last three decades have seen a growing number of studies that apply genealogies to quantitative questions concerning fertility, mortality, social mobility, and long-run inequality.

II.A Fertility and the Quantity–Quality Tradeoff

Lineage evidence from the late Ming and Qing shows that wealthy branches often produced more surviving children—"the rich get children" (Harrell et al., 1985). Yet a growing body of work identifies moderation in marital fertility and emerging child-quality investment. Shiue (2017) documents a quantity–quality trade-off in Tongcheng (Anhui). Using five Fujian and Guangdong lineages (about 50,000 individuals), Hu (2023) and Hu (2025) show that marital fertility in Ming–Qing China was moderate by global standards

[1] Before removing 337 individuals who bore the surname 黎 (a homophone of 李), the dataset contained 192,647 records.

and largely free of parity controls, challenging the view of universally high Chinese fertility.

II.B Mortality and Health Patterns

Chinese genealogies often record ages at death, enabling the reconstruction of mortality profiles. Lee, Feng, and Campbell (1994) analyzed the Aisin Gioro genealogy (the Qing imperial family) and found elevated infant mortality despite elite status, suggesting limits to privilege in a disease-laden environment. Zhao (2001) demonstrated that mortality derived from genealogical data can be severely biased upward due to survivorship selection—families that failed to reproduce or maintain the lineage leave no records. He proposed micro-simulation techniques to correct for this, showing that unadjusted life expectancies are significantly overstated.

II.C Elite Reproduction and Social Mobility

Because genealogies trace kinship over many generations, they allow researchers to study elite persistence and intergenerational mobility over long time horizons. Using digitized genealogies from Tongcheng, Shiue (2025) measures intergenerational mobility from the fourteenth through the nineteenth centuries. Her results show that upward mobility increased in the seventeenth century and declined thereafter, coinciding with shifts in local inequality. The analysis also documents fertility tradeoffs, as families with educated sons had fewer children. Other studies confirm that kinship networks and cultural capital helped sustain socioeconomic status across generations, even into the PRC era (Campbell and Lee, 2011).

II.D Occupational Structure

A small but significant subset of twentieth-century *jiapu*, compiled or revised after the 1980s, lists the occupation of every recorded adult—male and female alike. These “transformed” genealogies open a new window onto China’s modern labor structure, for which reliable census microdata exist only from 1982 onward. The Yangtze Jiapu Dataset assembled by Dai (2025) covers 210,383 occupational observations and permits the first direct estimates of sectoral shares for the late nineteenth century, 1933, and the reform era. By filling the pre-1982 gap, such data sharpen debates about structural change during the Republican and early PRC periods.

II.E Bias and Coverage Limitations

Despite their value, Chinese genealogies are highly selective sources. Most *jiapu* document surviving patrilines, with daughters and childless sons often omitted; as a result, naïve aggregates tend to understate mortality and overstate life expectancy, and can also overstate fertility for the lineages that persisted (Zhao, 2001). Geographic and social coverage is uneven: preserved volumes disproportionately come from wealthier lineages in the South and along the coast, limiting representativeness for Northern and interior regions. Within any lineage, main branches are typically better documented than cadet ones, and some compilations involve retrospective edits or reconstructions, which can introduce chronological inaccuracies. Modern digitization adds further challenges—duplicate entries, inconsistent toponyms, and linkage errors—unless carefully reconciled. These limitations warrant caution in generalizing to the population at large. Even so, the longitudinal depth and kin-linked structure of genealogies make them indispensable for studying the evolution of Chinese demographic and social systems when analyzed with appropriate corrections and validation.

II.F Contribution of This Study

This paper builds on the tradition of using genealogical sources and makes three contributions. First, it shows that the FamilySearch crowd-sourced sample is reasonable for historical demography, since descriptive patterns in population growth, sex ratios, and migration behave as expected and pass internal checks. Second, it provides a reproducible pipeline that turns FamilySearch into a longitudinal dataset for China through harvesting, de-duplication, CHGIS-based place standardization, and plausibility checks, demonstrated on one surname but easily extended to others. Third, it demonstrates analytical use at scale, with hypothesis-driven tests on sex ratios, migration, and kinship yielding consistent and interpretable relationships despite known biases.

III Methodology

Our empirical exercise proceeds in four steps: (i) harvesting surname–specific records from the *FamilySearch* genealogical tree; (ii) reconnecting those records into complete multigenerational pedigrees; and (iii) validating, de-duplicating, and geocoding the cleaned individuals. Each step is automated in `Stata` and `Python` scripts that can be re-run for any other surname.

III.A Data retrieval

The present paper illustrates the workflow with the surname “李” (*Li*). Relying on the application’s public search interface, we exported all entries whose primary surname field matches “李/Li”.[2] Data were retrieved using both Chinese-character and pinyin-based queries to ensure full surname coverage. Because the platform’s indexing is not always consistent across formats, this dual approach proved essential. Every downloaded record includes birth and (where available) death years, basic kin ties, and free-text place descriptions.

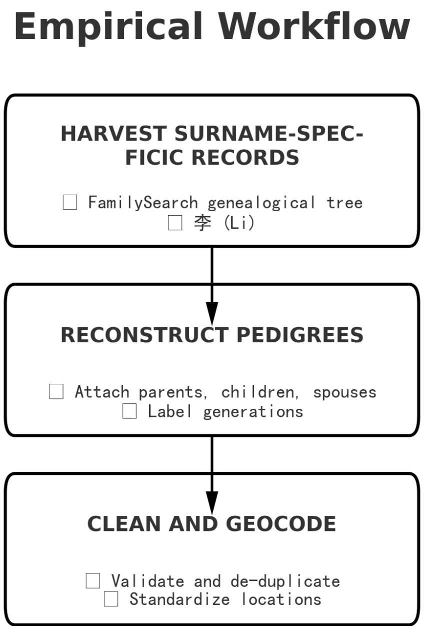

Empirical Workflow
HARVEST SURNAME-SPEC-FICIC RECORDS
□ FamilySearch genealogical tree
□ 李 (Li)
RECONSTRUCT PEDIGREES
□ Attach parents, children, spouses
□ Label generations
CLEAN AND GEOCODE
□ Validate and de-duplicate
□ Standardize locations

Figure I

Note: The diagram summarizes the preprocessing used in this paper: (i) harvest surname–specific records from a crowd-sourced genealogical tree (FamilySearch in this application; illustrated with 李/Li); (ii) reconnect individuals into multigenerational pedigrees by attaching parents, children, and spouses and labeling generations; and (iii) clean and geocode by de-duplicating records and standardizing locations (CHGIS). Construction of analysis-specific variables (e.g., population growth, sex ratios, migration distances) is documented in the relevant sections.

III.B Reconstructing pedigrees

Because *FamilySearch* indexes individuals one-by-one, the first task is to re-assemble those single nodes into coherent lineage trees. We iteratively attach every parent, child,

[2] The query returns every person—male *or* female—whose indexed surname is Li, including women who appear only as spouses in non-Li pedigrees but retain their natal surname in the database.

and spouse referenced in the data, stopping when no new IDs appear. Within each connected component, the most distant recorded ancestor is labeled generation 1, his children generation 2, and so on. Recursive backtracing is used to define each individual's "family root" as the ID of the earliest patrilineal ancestor whose own parent ID cannot be resolved; this identifier is then used to represent the entire lineage descended from him. The resulting file contains 535 distinct family roots, the deepest of which spans 19 generations.

III.C Cleaning and geocoding

Quality checks remove entries with clearly impossible information (e.g. lifespans exceeding 120 years without corroboration). Duplicate profiles created by alternate spellings are consolidated by comparing {surname, given name, province, birth year} keys and verifying that linked relatives match. Some individuals initially appeared only in relational fields (as children or spouses) and not in the main index; these were subsequently added through targeted ID-based queries.

Place strings are standardized using the FamilySearch place authority, which harmonizes historical and vernacular names to modern equivalents. Each record is then linked to county-level administrative units through the GB2000 coding system, supplemented by the China Historical GIS for geographic coordinates and boundary information. These standardized identifiers provide a consistent basis for subsequent spatial analysis.

IV Descriptive Patterns

This section examines three variables commonly used in demographic research: population growth, sex ratios, and migration. Each represents a basic dimension of population structure and change. We calculate these measures using the genealogical data as observed, without presupposing their completeness or accuracy. The purpose of this section is to describe what can be constructed from the data, and to establish a basis for later investigation into the sources of variation across regions, cohorts, and generations.

IV.A Population Growth

Because Chinese lineage records are overwhelmingly patrilineal, male lifespans are much more systematically recorded than female ones. Following standard practice in the genealogical-demography literature, our baseline series is constructed from men only. For completeness, we also provide the corresponding estimates for the full sample (men and women combined) in Appendix Figure A.1.

The provincial panel is built through a harmonized four-step routine, run separately for Hebei, Fujian, Guangdong, Hunan, and Zhejiang, with only the list of benchmark years varying by province. First, the data are restricted to males born in the target province, retaining only those with numeric and non-missing birth and death years. Second, for each benchmark year t we construct an indicator

$$\mathsf{alive}_{it} = 1\big(\text{birth}_i \leq t < \text{death}_i\big),$$

and sum across individuals to obtain the provincial head-count $N_{pt} = \sum_i \mathsf{alive}_{it}$. Third, any decade with fewer than 200 observed males is dropped to avoid spurious variation; this affects only a few province–decade cells in the early seventeenth and late nineteenth centuries. Finally, for all remaining cells we compute the *annualised* log growth rate

$$g_{pt} = \frac{\ln N_{pt} - \ln N_{p,t-10}}{10}, \qquad (t \geq 1610 \text{ and available in } p),$$

so that a value of $g_{pt} = 0.02$ corresponds to approximately 2% growth per year over the preceding decade.

Unique IDs in the cleaned database ensure no double-counting, and dropping cases with unknown death years guarantees that survivorship spells are fully observed. The resulting province-by-decade panel underlies Figure II and the regressions that follow.

The reconstructed male series traces the classic demographic arc: robust expansion through the seventeenth and eighteenth centuries; a dramatic mid-nineteenth-century collapse centred on the 1850s (coinciding with the Taiping civil war); and a partial rebound thereafter. Commercial Zhejiang and coastal Guangdong consistently outpace the aggregate, whereas Hebei lags, reflecting harsher climatic shocks and recurrent political turmoil in the north-China core. Equivalent results for the full sample, including women, are shown in Appendix Figure A.1.

IV.B Population Sex Ratio

To mitigate female undercoverage inherent in patrilineal genealogies, we construct the universe by appending the spouse file to the Li-surname individual file, then deduplicating on person ID so that each person appears once. We retain records with non-missing sex, birth year, and provincial birthplace, form province–century cells for the

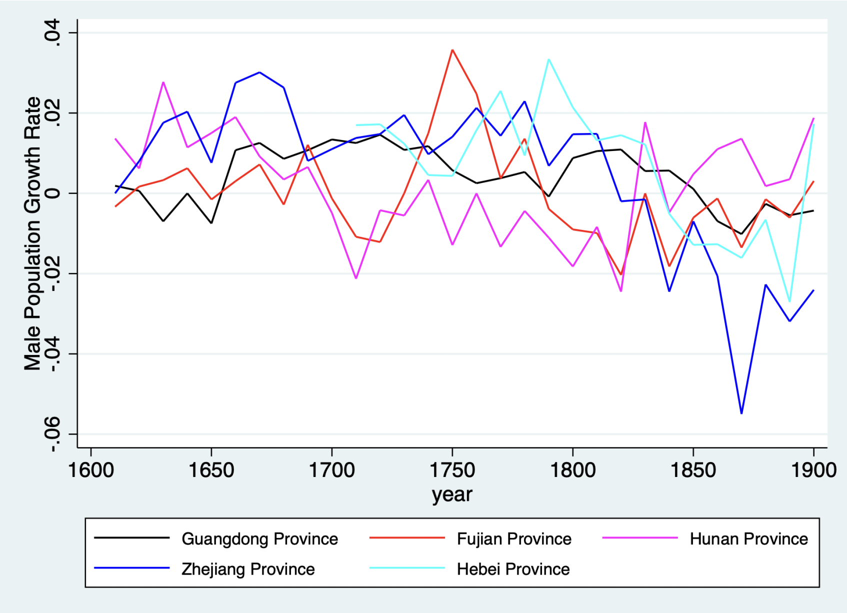


Figure II: Male population growth rates by decade, 1600–1900

Note: The y-axis reports annualized log growth, $g_{pt} = [\ln N_{pt} - \ln N_{p,t-10}]/10$. Thus $0.02 \approx 2\%$ per annum.

seventeenth through nineteenth centuries, and focus on large-sample provinces (Anhui, Fujian, Guangdong, Hebei, Henan, Hainan, Hunan, Jiangxi, Shandong, Zhejiang, and Taiwan). For each cell, the sex ratio is defined as males per 100 females.

Table I: Sex Ratios by Province and Century (17th–19th)

Province	17th	18th	19th	Average
Hebei	106.3	115.1	116.3	112.6
Shandong	104.8	96.0	106.2	102.3
Henan	127.8	132.3	129.0	129.7
Anhui	131.0	125.7	109.2	121.9
Jiangxi	102.1	99.5	98.8	100.1
Hunan	110.9	130.4	102.7	114.7
Zhejiang	90.2	121.6	117.1	109.6
Fujian	159.1	167.2	168.3	164.9
Guangdong	114.7	107.1	110.0	110.6
Hainan	110.9	109.9	103.8	108.2
Taiwan	126.8	129.5	114.1	123.5

Note: Sex ratio is males per 100 females. Universe combines Li-surname individuals and their spouses after de-duplication. Provinces are assigned by birthplace. Centuries follow birth year groups. Tibet excluded due to small counts.

Including spouses yields patterns that are plausible and closer to population composition. Averaged over the seventeenth–nineteenth centuries, Fujian is the most male-biased at about 165 males per 100 females; Henan is around 130; Taiwan about 124; Guangdong and Zhejiang are roughly 111 and 110; Hebei and Hunan are near 113 and 115; Hainan is about 108; Shandong is close to 102; and Jiangxi is essentially at parity at about 100. These cross-province gradients on the 17th–19th century averages reinforce that the combined individual-plus-spouse construction yields sensible sex-ratio profiles.

IV.C Extent of Migration

We define an individual as an out-migrant if the recorded place of death lies outside the administrative polygon of the parent's origin (the parent's death place when available, or birthplace otherwise). If the two polygons are nested—for example, a county within its prefecture—the move indicator and any associated distance measure are set to zero.

Applying this rule requires several restrictions. The analysis is limited to 37,375 parent–child pairs where both parties' locations can be matched to county-level polygons or to clearly identified foreign localities such as "Singapore" or "Penang." A further requirement is that the child can be linked to at least one identifiable parent with a known location. Records lacking such a link—overwhelmingly women who appear only as spouses in their husbands' genealogies—are excluded. Because women are much more likely to lack a parental link, they constitute only about three per cent of the geo-coded sample and fewer than one per cent of detected moves. The statistics that follow therefore primarily reflect male mobility. Entries with unmatched or overly vague locations (e.g. records listing only "China") are excluded from the migration analysis.

Moves to destinations outside China (e.g. the Straits Settlements or Siam) enter the migration counts and regressions but are omitted from distance-based calculations and plots. Among the 37,375 geo-coded parent–child pairs, we identify 894 moves, an out-migration rate of 2.4 per cent. Figure III disaggregates these rates by birth century and lineage.

For each move we calculate the great-circle distance between the centroids of origin and destination polygons, using two alternative definitions of origin—the parent's place of death or, if missing, the parent's birthplace. Moves between nested polygons or with centroids less than 100 m apart are assigned a distance of zero,[3] so minor errors do not

[3] The 100 m tolerance and the parent–child nesting rule prevent minor boundary misalignments and geocoding noise from inflating the count of long-distance moves.

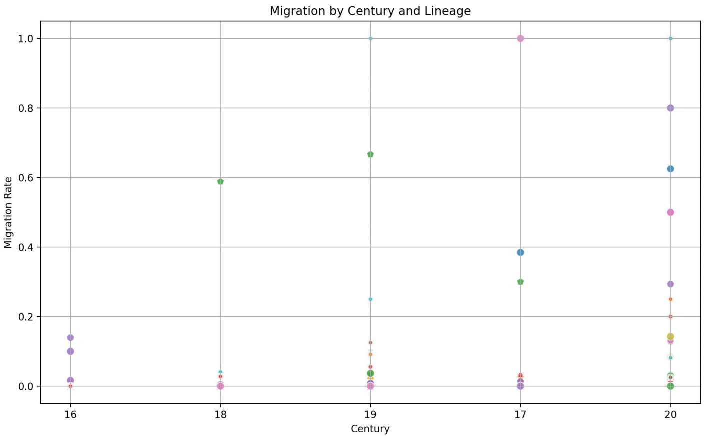
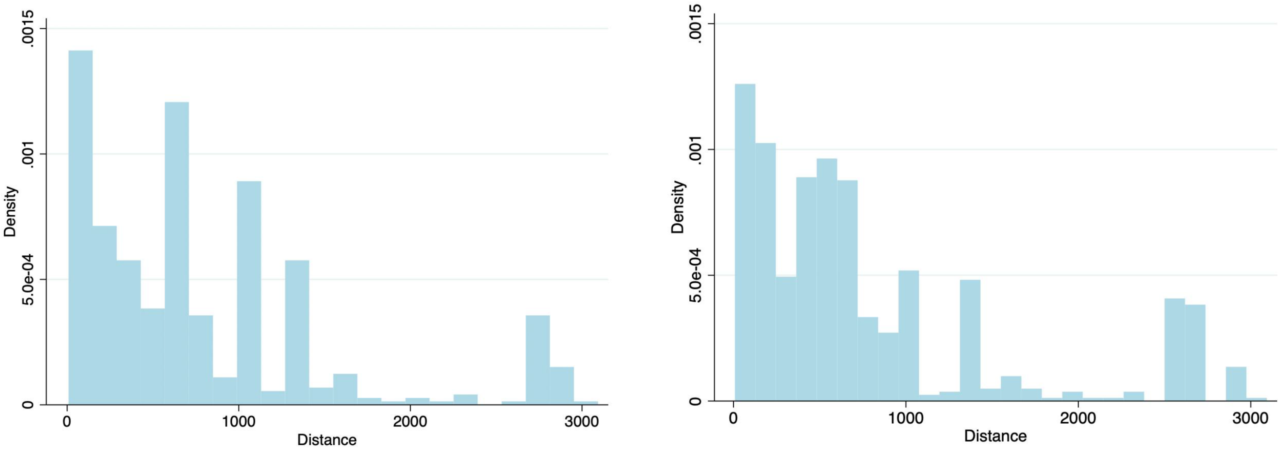


Figure III: Out-migration rate by century and lineage

Note: The plot is restricted to the ten lineages that meet two stability criteria: (i) at least 500 geo-coded individuals in total and (ii) a minimum of ten observations in every represented generation. Each marker reports the share of persons in a lineage–century cell whose place of death lies outside that of their parent. Marker area is proportional to the number of geo-coded parent–child pairs in the cell, and color identifies the lineage. Although no gender filter is imposed, fewer than five per cent of records are female, so the graphic largely reflects male moves.



(a)

(b)

Figure IV: Histogram of Migration Distances Inferred from Parent–Child Pairs

Note: Panels plot great-circle distances (km) for all moves with measurable origin–destination pairs. The left panel uses the parent's place of death as origin; the right panel uses the parent's birthplace. Moves to overseas destinations and records with only country-level locations (e.g. "China") are excluded from these distance plots.

inflate mobility measures. The resulting distribution (Figure IV) is highly skewed. More than half of moves are under 50 km, while a thin tail extends beyond 1,500 km. These long-distance moves are infrequent and heterogeneous and do not concentrate in any single origin–destination pattern.

Taken together, the evidence points to an overwhelmingly sedentary population: fewer than three per cent of parent–child pairs involve an out-migration, and more than half of those moves cover under 50 km. A small subset of lineages nevertheless displays markedly higher mobility—sometimes spanning hundreds of kilometers and, in rare cases, crossing national borders—highlighting pronounced heterogeneity behind the overall low migration rate.

V Testing Demographic Hypotheses with Genealogical Data

V.A Skewed Sex Ratios and Male Childlessness

We test the hypothesis that male-biased sex ratios reduce male reproductive success. The outcome is the probability that a man has at least one recorded child, where childlessness is defined as the absence of linked children in the genealogy. Because our interest is in male childlessness, we restrict the sample to men.

Table II: Sex Ratio and Probability of Having Any Recorded Child

	(1)	(2)	(3)	(4)
Dependent variable: Child Dummy	OLS	Logit	OLS	Logit
Sex Ratio	-0.00136^{***}	-0.00548^{***}	-0.00171^{***}	-0.00698^{***}
	(0.000424)	(0.00172)	(0.000305)	(0.00125)
Century FE	No	No	Yes	Yes
Observations	95,613	95,613	95,613	95,613
R^2/Pseudo R^2	0.0021	0.0015	0.016	0.011

Note: The table reports estimates of the relationship between province–century sex ratios (males per 100 females) and the probability that a man has at least one recorded child. Columns 1 and 3 report OLS regressions, while columns 2 and 4 report logit estimates. Columns 3 and 4 include century fixed effects. The dependent variable equals one if the individual has at least one recorded child. The sample excludes women and those born after 1900. Sex ratios are calculated at the province–century level from the Li-surname universe augmented with spouses after de-duplication. Robust standard errors clustered by province of birth in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Sex ratios are defined at the province–century level from the Li-surname universe augmented with spouses after de-duplication, as described in Section IV.B. Each value represents the number of males per 100 females with known sex and provincial birthplace.

We regress the child dummy on the province–century sex ratio. Columns 1 and 2 of Table II present simple OLS and logit specifications with province-clustered robust standard errors. Columns 3–4 add century fixed effects, so coefficients are identified from within-century deviations of each province–century sex ratio from the century mean.

As shown in Table II, the results consistently indicate a negative and statistically significant relationship: in all specifications, higher male-to-female ratios are associated with a lower likelihood that a man is recorded as having children. In the fixed-effects OLS specification, a one-point increase in the sex ratio is associated with a 0.171 percentage-point decrease in the probability of recorded offspring. The corresponding logit specification confirms the same pattern, with similar magnitudes. We view these results as strong descriptive evidence that demographic competition among men, driven by skewed sex ratios, limited reproductive opportunities.

V.B Lineage and Mobility

Scholars since Freedman (1958, 1966), and more recently Watson (1988) and Szonyi (2002), argue that Chinese patrilineal descent groups pass through a characteristic development cycle. Founders are portrayed as mobile pioneers who leave their native districts to claim land or to escape adverse political or ecological shocks. Once a viable settlement is in place, however, the very institutions that sustain the group—ancestral halls, graveyards, collaborative landholdings, and an active ancestral cult—anchor descendants to the locality. Following the South China lineage literature, we use *lineage* to denote this localized, genealogy-anchored corporation.[4] A straightforward implication is high migration in the first few generations followed by progressive consolidation and spatial inertia.

We test this hypothesis with the genealogical sample. Migration is coded at the individual level: a male is an out-migrant if the county- or prefecture-level polygon of his death place lies outside that of his father.[5] Generations are enumerated within each `family_root`, with generation 1 denoting the recorded ancestor at the top of the genealogy.

Figure V visualizes the broad pattern. Founding generations display extreme heterogeneity—some lineages remain sedentary, others are entirely migrant—but the en-

[4] Much of the broader literature uses "clan" for similar entities. Our empirical unit corresponds to what anthropologists term a lineage, and we treat "clan" in cited work as synonymous unless scale distinctions matter.

[5] The coding rules follow the procedure introduced in Section IV.C. Results are virtually identical if we use paternal birth rather than death polygons.

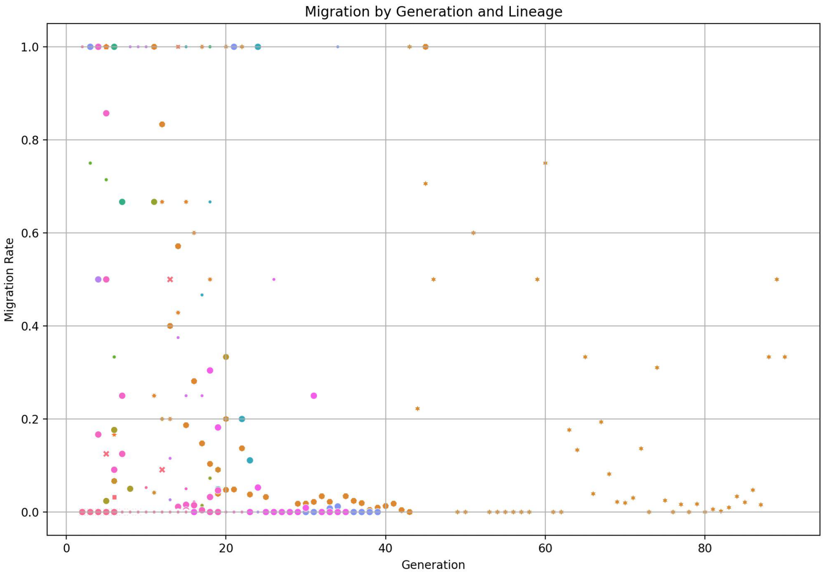


Figure V: Out-Migration Rate by Generation and Lineage

Note: Each point represents the migration rate of a patrilineal generation within a specific family root. The vertical axis plots the share of male individuals in a given generation who are classified as out-migrants, based on a mismatch between their place of death and that of their father. Generations are numbered from the founding ancestor recorded in each lineage. The variation in marker color and size reflects different family roots and sample sizes, respectively. Migration rates are highly variable in early generations and tend to cluster near zero in later generations, though occasional spikes appear among smaller lineages.

Table III: Effect of Generation on the Probability of Out-Migration

	Dependent variable: Out-Migration (0/1)
VARIABLES	(1)
Generation	-0.0002193^{***}
	(3.76×10^{-6})
Time (50-yr cohort) FE	Yes
Observations	29,615
R-squared	0.147

Note: HDFE regression absorbing 50-year birth-cohort dummies. Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

velope of rates narrows quickly. By generation 20, virtually every lineage records fewer than five percent out-migrants; many record none. Appendix Figure A.2 traces the ten largest lineages in our geo-coded sample: migration rates for nine of the ten taper to near zero by about generation 15, whereas one outlier lineage continues to send movers well into later generations.

Table III confirms the descriptive trend. Regressing the binary migration indicator on patrilineal generation and absorbing 50-year birth-cohort fixed effects yields a highly significant negative slope: each additional generation lowers the probability of out-migration by about 0.022 percentage points. Adding lineage-level controls (number of spouses, children, or total branch size; not shown) does not alter the estimate.

Taken together, the genealogical evidence lends empirical support to the settling-in narrative: mobility is a distinctive feature of the founding generations but fades as the lineage embeds itself in local ritual, economic, and political networks. Occasional late-generation spikes remind us that external shocks—wars, famine, or state resettlement programs—could still break the grip of locality (Shiue and Keller, 2024), yet these are exceptions rather than the rule. The broader implication is that Chinese lineages were simultaneously engines of geographic expansion and mechanisms of spatial fixity: once roots took hold, the very institutions that enabled collective power also curtailed further movement.

VI Conclusion and Future Directions

Using more than 190,000 *FamilySearch* records for a single common surname, this study shows that Chinese genealogical data can be converted—via systematic deduplication, CHGIS-based place standardization, and plausibility checks—into a large, internally coherent panel suitable for quantitative analysis. Descriptive patterns in the cleaned panel behave as expected and serve as internal validation. The male population series traces seventeenth–eighteenth-century expansion, a sharp mid-nineteenth-century contraction, and partial recovery. Sex ratios are markedly male-biased in several provinces—especially in the South (e.g., Fujian)—but pronounced imbalances also appear in parts of the North (e.g., Henan), with near-parity in others. Migration is overwhelmingly local, with a thin, heterogeneous long-distance tail.

Beyond these diagnostics, the hypothesis-testing exercises illustrate what the dataset enables. Results show that marriage-market imbalance aligns with a higher share of men who leave no recorded offspring and that deeper kinship depth is associated with lower out-migration. We read these as demonstrations of use rather than headline causal claims, highlighting how a cleaned *FamilySearch* panel can support targeted, hypothesis-driven analyses in historical demography.

A central virtue of the *FamilySearch* approach is that it complements—and in many cases out-scales—the traditional strategy of digitizing complete genealogies one lineage at a time. Full, cover-to-cover editions of individual clan books remain indispensable for studying ritual practice, property transfers, or the evolution of a single lineage's internal rules. Yet compiling such editions is labor-intensive and inherently piecemeal: each new clan requires a separate archival quest, specialized transcription, and bespoke coding. By contrast, the *FamilySearch* tree aggregates material across thousands of clans simultaneously and exposes it through a uniform application interface. With the cleaning pipeline presented here, researchers can harness that breadth to extract macro-demographic indicators—population growth, sex ratios, migration flows—on a scale that would be prohibitively costly if every genealogy had to be digitized clan-by-clan.

The pipeline is readily extensible. Replicating it for other common surnames would raise the sample into the millions and provide the scale necessary for large-scale analyses of Chinese historical demography. FamilySearch currently contains over two million Chinese records of similar structure, making it possible to apply the same procedures far beyond the single surname examined here. Expanding coverage in this way would strengthen the statistical basis for detecting long-run demographic patterns and regional variation. Automating the recovery of women and collateral branches—groups often under-recorded in traditional lineage books—remains a priority for reducing gender bias. Finally, the prominent Guangdong-to-Southeast-Asia corridor already visible in the pilot suggests a rich agenda on how emigrant communities transplanted or adapted lineage practices overseas. By combining multiple surnames and scaling up the analysis, researchers can build truly large historical microdatasets and revisit enduring questions about family, migration, and economic change in late imperial and modern China.

References

Buckles, Kasey, Adrian Haws, Joseph Price, and Haley E. B. Wilbert. 2023. *Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project.* NBER Working Paper 31671. National Bureau of Economic Research.

Buckles, Kasey, Joseph Price, Zachary Ward, and Haley E. B. Wilbert. 2023. *Family Trees and Falling Apples: Historical Intergenerational Mobility Estimates for Women and Men.* NBER Working Paper 31918. National Bureau of Economic Research.

Campbell, Cameron, and James Z Lee. 2011. "Kinship and the Long-Term Persistence of Inequality in Liaoning, China, 1749-2005." *Chinese Sociological Review* 44 (1): 71–103.

Colasurdo, Andrea, and Riccardo Omenti. 2024. "Using Online Genealogical Data for Demographic Research: An Empirical Examination of the FamiLinx Database." *Demographic Research* 51 (41): 1299–1350.

Dai, Ying. 2025. "Lineage Genealogies as A New Source for Researching the Occupational Structure of Twentieth-Century China: Tradition (Partially) Transformed." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 58 (1): 54–79.

Freedman, Maurice. 1958. *Lineage Organization in Southeastern China.* Published in association with the London School of Economics and Political Science. London: University of London, Athlone Press.

———. 1966. *Chinese Lineage and Society: Fukien and Kwangtung.* London: Athlone Press.

Harrell, Stevan, et al. 1985. "The Rich Get Children: Segmentation, Stratification, and Population in Three Chekiang Lineages, 1550-1850." *Family and population in East Asian history,* 81–109.

Hu, Sijie. 2023. "Descendants over 300 Years: Marital Fertility in Five Lineages in Qing China." *Asia-Pacific Economic History Review* 63 (2): 200–224.

———. 2025. "Evolutionary Advantage of Moderate Fertility during Ming–Qing China: A Unified-Growth Perspective." Forthcoming, *Journal of Economic Growth.*

Kaplanis, Joanna, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Michael Wahl, Maya Gershovits, et al. 2018. "Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives." *Science* 360 (6385): 171–175.

Lee, James, Wang Feng, and Cameron Campbell. 1994. "Infant and Child Mortality Among the Qing Nobility: Implications for Two Types of Positive Check." *Population Studies* 48 (3): 395–411.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. 2021. "Combining family history and machine learning to link historical records: The Census Tree data set." *Explorations in Economic History* 80:101391.

Shiue, Carol H. 2025. "Social Mobility in the Long Run: A Temporal Analysis of Tongcheng, China, 1300 to 1900." *Journal of Economic History.*

Shiue, Carol H, and Wolfgang Keller. 2024. *Elite Strategies for Big Shocks: The Case of the Fall of the Ming.* Technical report. National Bureau of Economic Research.

Shiue, Carol H. 2017. "Human Capital and Fertility in Chinese Clans Before Modern Growth." *Journal of Economic Growth* 22, no. 4 (December): 351–396.

Szonyi, Michael. 2002. *Practicing Kinship: Lineage and Descent in Late Imperial China.* Stanford University Press.

Watson, Rubie S. 1988. "Remembering the Dead: Graves and Politics in Southeastern China." In *Death Ritual in Late Imperial and Modern China,* edited by James L. Watson and Rubie S. Watson, 121–143. Berkeley: University of California Press.

Zhao, Zhongwei. 2001. "Chinese Genealogies as a Source for Demographic Research: A Further Assessment of Their Reliability and Biases." *Population Studies* 55 (2): 181–193. http://www.jstor.org/stable/3092962.

Appendix

Additional Figures and Tables

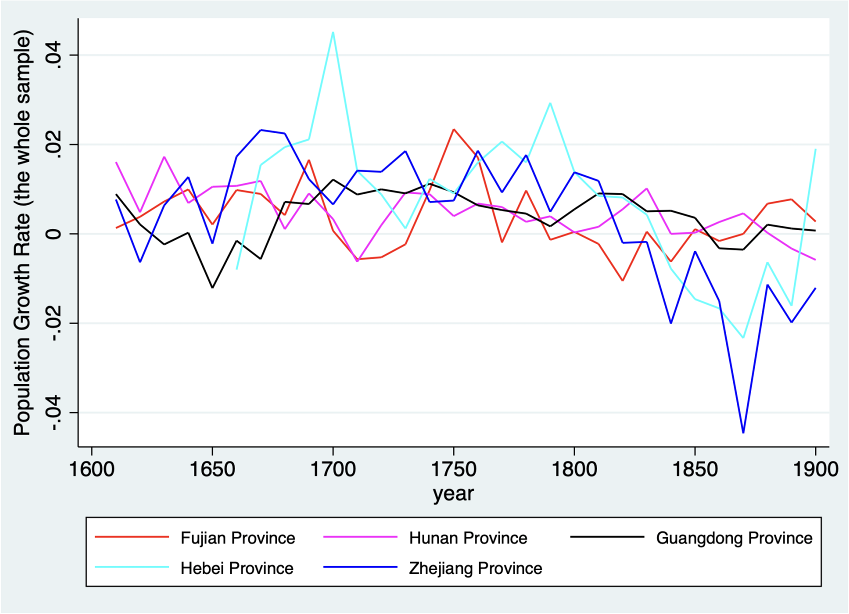


Figure A.1: Population growth rates by decade, 1600–1900

Note: y-axis shows annualised log growth, $g_{pt} = [\ln N_{pt} - \ln N_{p,t-10}]/10$; thus $0.02 \approx 2\%$ per annum.

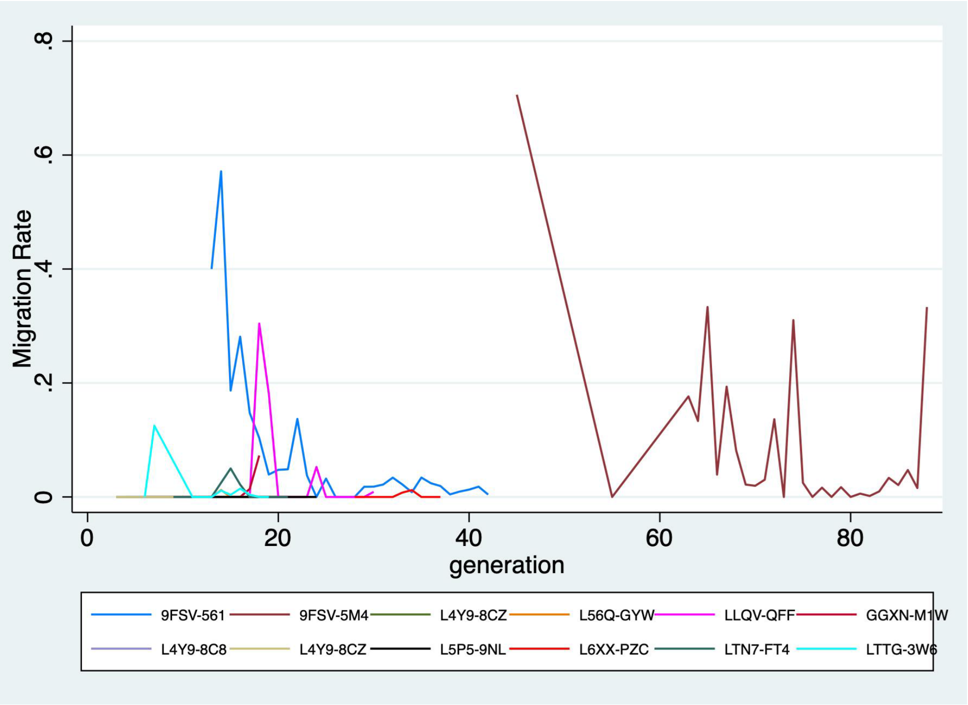


Figure A.2: Generation-by-Generation Migration Rates for Ten Large lineages

Note: Each colored line tracks the proportion of sons whose recorded place of death lies outside their fathers' origin polygon, plotted by patrilineal generation. The ten lineages shown are those with at least 200 geo-coded father–son pairs; line segments are suppressed when a generation contains fewer than five observations. The figure highlights the rapid settling-in that follows a lineage's initial move, alongside the rare but persistent mobility of one outlier clan whose migratory activity extends beyond the 80th generation.