Explorations in Economic History 99 (2026) 101734

Contents lists available at ScienceDirect

Explorations in Economic History

journal homepage: www.elsevier.com/locate/eeh

Shorter Article

Crowd-sourced Chinese genealogies as data for demographic and economic history[☆]

Melanie Meng Xue[1]

London School of Economics and Political Science (LSE), United Kingdom

ARTICLE INFO

JEL classification:
J11
J13
N1
N35

Keywords:
Crowd-sourced genealogies
China
Migration
Sex ratios

ABSTRACT

This paper evaluates the usefulness of crowd-sourced Chinese genealogical data for quantitative research in demography and economic history. I first examine whether genealogies — despite well-known selection biases — produce demographic patterns consistent with established historical knowledge of China. Comparisons with existing studies show that aggregate population-growth trends and sex ratios over time align reasonably well with established demographic and historical findings, suggesting that genealogies, though selective, capture coherent and interpretable patterns. Building on these plausibility checks, the paper argues that the main value of genealogical data lies in their scalability and temporal depth, particularly as crowd-sourced digitization vastly expands the number of available records. These features make genealogies well suited to analyses that leverage variation across regions and over time, an approach that is central in modern economic history.

1. 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 availability of these data creates new opportunities for historical demography, but it also raises practical questions about validity and use. Genealogies have well-known selection biases and are not designed to represent entire populations. Yet if their patterns are broadly consistent with established demographic knowledge, they may still serve as a valuable basis for large-scale empirical research.

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

[☆] I thank the editor and two anonymous referees for their constructive feedback. I am especially grateful to Carol Shiue and Wolfgang Keller for their extensive and insightful comments. I also thank Cameron Campbell and Patrick Wallis for helpful comments, as well as a colleague for thoughtful suggestions that clarified the argument and improved the exposition. I thank Yan Li and Yuyang Zhang for excellent research assistance and 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 explored.

E-mail address: m.m.xue@lse.ac.uk.

[1] Assistant Professor at London School of Economics and Political Science (LSE), Visiting Fellow at Yale University, and Research Affiliate at the Centre for Economic Policy Research (CEPR).

https://doi.org/10.1016/j.eeh.2025.101734
Received 28 September 2025; Received in revised form 4 November 2025; Accepted 20 November 2025

Available online 29 November 2025

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 et al., 2023a,b). Demographers have also evaluated data quality and selection in aggregated online trees such as FamiLinx, identifying systematic survivorship and reporting biases and proposing correction methods (Chong et al., 2022; Stelter and Alburez-Gutierrez, 2022; Calderón-Bernal et al., 2023; Colasurdo and Omenti, 2024). Building on this international literature, this paper evaluates whether *crowd-sourced* Chinese genealogies can reproduce key demographic patterns and explores their broader analytical potential for quantitative research.

The paper focuses on one surname — 李 (*Li*) — to illustrate both the strengths and limits of the approach. Li is among the most common surnames in China, accounting 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 validation rules to flag likely errors. The resulting dataset contains 192,310 unique individuals spanning fifteen centuries and all provinces of China.[2]

Although the evidence comes from only one patriline, it aligns closely with established demographic knowledge. The population growth trajectory and shifts over time correspond to known historical trends; the sex ratios exhibit the familiar southern bias consistent with earlier demographic studies. These parallels suggest that, even with inherent biases, genealogical data contain coherent and interpretable information. Having established this consistency, the paper turns to the broader question of use: how such data can support the kinds of large-scale, regression-based analyses common in economics and economic history. The emphasis is not on reconstructing complete populations, but on assessing whether genealogical records can generate credible estimates and relationships across time and place.

By combining plausibility checks against established benchmarks with proof-of-concept analyses, the study lowers the entry cost for economists, demographers, and historians interested in China's long-run population dynamics. It shows that crowd-sourced genealogical data are a scalable input for quantitative research—complementing rather than replacing the more detailed but smaller-scope studies that have long defined Chinese historical demography.

2. Genealogical sources in Chinese demographic and economic history

This section reviews prior work that has used genealogical sources in Chinese economic history, organized by research themes including fertility, mortality, social mobility, and occupational structure.

Chinese genealogies (族谱, 宗谱) have long been used in historical demography and economic history, offering uniquely detailed evidence on family behavior, lineage organization, and social mobility. Their modern use began with anthropological and qualitative studies of kinship and ritual (Freedman, 1958, 1966), but quantitative applications expanded as larger collections became accessible. Early quantitative work by Liu (1978), Telford (1986), Harrell (1987), and Zhao (1994) explored how genealogies could be used to infer fertility, mortality, and lineage persistence, while also identifying the selection and survivorship biases inherent in such records. Because surviving volumes were few and geographically concentrated, most analyses focused on single clans or counties, reflecting both the goals of historical demography — reconstructing population-level parameters — and the severe limits of data access and computational power before large-scale digitization.

2.1. 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 rising investment in child quality. Shiue (2017) documents a quantity–quality trade-off in Tongcheng (Anhui), using data originally transcribed by Ted Telford. Using five Fujian and Guangdong lineages (about 50,000 individuals), Hu (2023), Hu find that marital fertility in Ming–Qing China was moderate by global standards and largely free of parity controls, challenging the view of universally high Chinese fertility; see also Clark (2007) for a broader comparative discussion of low pre-industrial fertility in East Asia.

2.2. Mortality and health patterns

Because many genealogies record ages at death, they allow approximate reconstruction of mortality profiles. Lee et al. (1994) analyzed the Aisin Gioro genealogy of 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 biased upward due to survivorship selection—families that failed to reproduce or maintain the lineage leave no records. He proposed microsimulation techniques to correct this bias, showing that unadjusted life expectancies are significantly overstated. These studies together clarified both the potential and the limits of genealogies for mortality analysis.

[2] Before removing 337 individuals who bore the surname 黎 (a homophone of 李), the dataset contained 192,647 records. A small number of records list overseas places of birth or death (for example Singapore, Penang, San Francisco); these are retained and coded as foreign locations in the migration analysis in Section 4.3.

2.3. 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. Other studies confirm that kinship networks and cultural capital helped sustain socioeconomic status across generations, even into the PRC era (Campbell and Lee, 2011). These findings connect the micro-structure of lineages to broader questions about the persistence of elites in China's long-run development.

2.4. Occupational structure

A subset of twentieth-century *jiapu*, compiled or revised after the 1980s, records the occupation of every listed adult—male and female alike. These modern compilations provide the first systematic view of China's labor structure before the advent of comprehensive census microdata. The Yangtze Jiapu Dataset assembled by Dai (2025) covers 210,383 occupational observations and permits 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 and show how genealogies can extend beyond demographic reconstruction to economic history.

2.5. Bias and coverage limitations

Despite their value, Chinese genealogies are highly selective sources. Most *jiapu* document only surviving patrilines, with daughters and childless sons often omitted. Naïve aggregates therefore tend to understate mortality and overstate life expectancy, and can also overstate fertility for lineages that persisted (Zhao, 2001). Zhao shows through microsimulation that these selection mechanisms materially bias level estimates derived from genealogies and proposes corrections that recover more plausible demographic quantities (Zhao, 1994, 2001).

Coverage is uneven in space and social status. Preserved volumes disproportionately come from wealthier southern and coastal lineages, which limits representativeness for northern and interior regions (Liu, 1978; Harrell, 1987). Within lineages, main branches are usually better documented than cadet ones, and some compilations include retrospective edits that can introduce chronological inconsistencies.

Direct linkage studies make these issues concrete. In Liaoning, Campbell and Lee (2002) link genealogies to household registers and show that short-lived and low-status individuals are more likely to be missing in genealogies, and that recorded populations diverge systematically from official registers. These findings reinforce the need to treat genealogies as incomplete population records.

2.6. Contribution of this study

This paper builds on this research tradition but with a different emphasis. The first objective is to evaluate whether newly digitized and crowd-sourced genealogical data from FamilySearch yield demographic patterns consistent with established knowledge. I focus on descriptive indicators — population growth, sex ratios, and migration — to assess how far these data replicate known trends and behave sensibly under internal checks. The second objective is to demonstrate that, after establishing their basic demographic plausibility, genealogical data can inform a broader range of analyses in economics and economic history. Rather than reconstructing population-level parameters, the aim is to show that these sources — despite selection and coverage biases — can generate credible estimates in regression-based research designs. This approach complements earlier demographic work by showing that the growing universe of digitized, crowd-sourced genealogies provides a scalable empirical foundation for studying long-run processes in China's economic and social history.

3. Data and methodology

Our empirical exercise proceeds in three 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 implemented through a standardized workflow that can be replicated for any other surname.[3]

Data origin and digitization. *FamilySearch* is a non-profit genealogical archive operated by the Church of Jesus Christ of Latter-day Saints (LDS Church). Its holdings on Chinese lineages fall into two categories: (a) scanned images of printed or manuscript *zupu* volumes acquired through library microfilming programs since the 1960s, and (b) structured genealogical records that have been digitized and indexed into the open-access *Family Tree* database. The present study draws exclusively on the latter: records already transcribed and linked through the Family Tree interface. These digitized entries are generated through a mixture of LDS–sponsored indexing campaigns and volunteer contributions by users in China and abroad, including — but not limited to — Church members.[4]

[3] For a field overview of Chinese genealogies, including their typical contents, known selection issues, and major repositories, see Shiue (2016).

[4] FamilySearch's Chinese holdings comprise both unindexed images of lineage books and a smaller, already-indexed subset searchable through the Family Tree; the latter is produced through a mix of computer extraction and global volunteer review via the Get Involved indexing program, which is not limited to Church members.

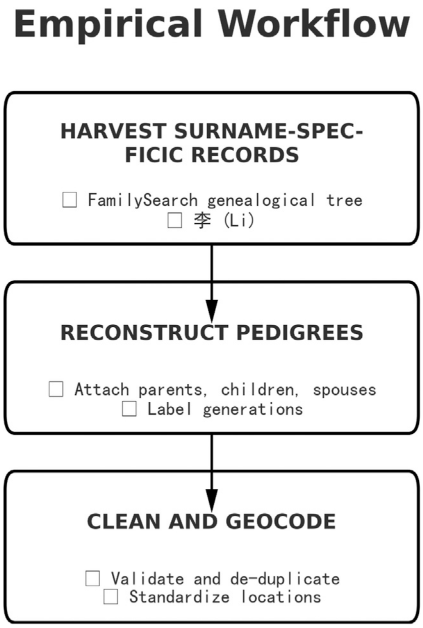
Empirical Workflow
HARVEST SURNAME-SPEC-FICIC RECORDS
□ FamilySearch genealogical tree
□ 李 (Li)
RECONSTRUCT PEDIGREES
□ Attach parents, children, spouses
□ Label generations
CLEAN AND GEOCODE
□ Validate and de-duplicate
□ Standardize locations

Fig. 1.
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. See also Appendix Figure A.1 for an illustrative example of the FamilySearch person page from which these structured fields and relationship links are obtained.

Each record reflects a combination of automated optical character recognition (OCR), manual data entry, and ongoing crowd-sourced editing. As of 2025, only a small fraction of the Chinese genealogies in FamilySearch's image archive have been converted into structured, searchable data, but even this subset represents millions of person-level entries spanning nearly all provinces. Our workflow operates directly on this public, structured component; we do not use unpublished image-only volumes.

As illustrated in Figure A.1, each person record on *FamilySearch* is a node with a unique Person ID (PID), and the interface lists linked relatives (parents, siblings, spouses, children). Our raw download captures these PIDs and the relationship links, which we later formalize as graph edges.

3.1. 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".[5] 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 (see Fig. 1).

Figure A.1 shows the person page format we harvest: a PID-identified node (e.g., `G97R-BT8`) and the full set of linked PIDs for spouses, parents, and children. These fields constitute the inputs to the reconstruction step.

3.2. Reconstructing pedigrees

We treat the FamilySearch links (Fig. A.1) as a multiplex graph on the same node set: a directed parent→child layer and an undirected spouse layer. Starting from all Li-surname PIDs, we iteratively attach every linked PID in both layers until no new nodes appear, then assign generation labels within each connected component.

Because *FamilySearch* indexes individuals one by one, the first task is to reassemble those single nodes into coherent lineage trees. Each individual in *FamilySearch* has a persistent alphanumeric *person ID* that uniquely identifies the record and links parents,

[5] 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.

spouses, and children across entries. We iteratively attach every parent, child, 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 *family root* identifier serves as a unique lineage key that represents the entire set of descendants from that ancestor. The resulting file contains 535 distinct family roots.

3.3. 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.

Place strings are standardized using the FamilySearch place authority. Each record is then linked to county-level administrative units using China's national administrative-division codes (GB/T 2260; hereafter "GB2000") (Standardization Administration of China, 2013), and supplemented with geographic coordinates and historical boundaries from the China Historical GIS (Fairbank Center for Chinese Studies and the Institute for Chinese Historical Geography at Fudan University, 2012). These standardized identifiers provide a consistent basis for subsequent spatial analysis.

4. Demographic patterns and comparison with historical evidence

This section presents three basic indicators — population growth, sex ratios, and migration — constructed from the genealogical data and, where possible, compares them with established findings in Chinese historical demography. The aim is not to reproduce population-wide levels but to document what can be constructed from the data and to assess whether the resulting patterns are historically plausible.

4.1. Population growth

Because Chinese lineage records are overwhelmingly patrilineal, we construct a male-only baseline series; the combined series (men and women) is shown in Appendix Figure A.2.

The provincial panel is built via a harmonized routine, run separately for Hebei, Fujian, Guangdong, Hunan, and Zhejiang: (i) restrict to males born in the target province with numeric, non-missing birth and death years; (ii) for each benchmark year t, construct $\mathsf{alive}_{it} = \mathbb{1}(\text{birth}_i \le t < \text{death}_i)$ and sum $N_{pt} = \sum_i \mathsf{alive}_{it}$; (iii) drop decades with fewer than 200 observed males; (iv) compute the annualized log growth rate

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

so that $g_{pt} = 0.01$ corresponds to about 1% per annum 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 Fig. 2 and the analyses that follow.

The reconstructed male series traces the familiar demographic arc: robust expansion through the seventeenth and eighteenth centuries, a dramatic mid-nineteenth-century collapse centered on the 1850s (coinciding with the Taiping civil war), and a partial rebound thereafter. Patterns are broadly similar across provinces. Equivalent results for the full sample, including women, are shown in Appendix Figure A.2.

Benchmark comparison. We compare *genealogical male headcount growth* in Fig. 2 to *national population growth* benchmarks in Appendix Table A.2 (A1), which imply roughly 0.47% per annum for 1776–1820 and 0.44–0.49% per annum for 1820–1851; turning points align closely across series. As a simple level check based on external fertility, we take the Hsiao-shan (Zhejiang) first-wife TFRs in Appendix Table A.3 and map them to implied intrinsic growth via

$$R_0 \approx \text{TFR} \times p_f \times s, \qquad r \approx \frac{\ln R_0}{T},$$

using $p_f \approx 0.488$ (SRB ≈ 105), $s \in [0.45, 0.50]$ (share of daughters surviving to mean age at childbearing), and $T \approx 27$ years. Appendix (A3) reports the resulting range. For mid-eighteenth-century anchors TFR = 5.06–5.29, the implied r is about 0.39–0.95% per annum; across the full Hsiao-shan set TFR = 4.63–6.38, the implied band spans roughly 0.06–1.64% per annum (A3). Read against these calibrated bands, the Zhejiang line in Fig. 2 is of the same order of magnitude in the relevant decades; point estimates are slightly higher, consistent with (i) the anchor's *first-wife only* fertility being conservative relative to total marital fertility, and (ii) reasonable variation in T and s.[6]

[6] Results are similar for $T \in [26, 28]$ and $s \in [0.40, 0.55]$. The first-wife restriction in Appendix Table A.3 understates total births per father, so implied growth bands are conservative.

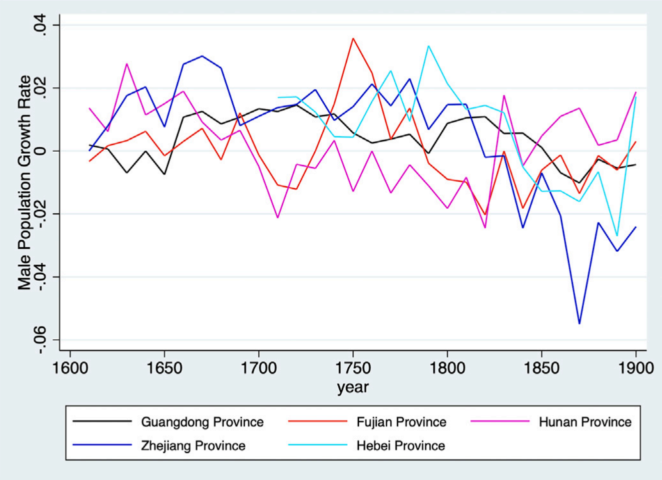

Fig. 2. Male headcount growth by decade from genealogies, 1600–1900.
Note: $g_{pt} = [\ln N_{pt} - \ln N_{p,t-10}]/10$; Series reflect *genealogical male headcount growth*, not total population growth.

Table 1
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.

Cohort-specific caveat. For birth cohorts in the 1810s–1820s, the intrinsic rates implied by the Hsiao-shan fertility anchors exceed the *genealogical male headcount* growth in our series. This is expected. The intrinsic rate r is constructed from fertility and daughter survival to mean childbearing and therefore abstracts from mid- and late-life mortality shocks. By contrast, the headcount series cumulates deaths at all ages. Men born in the 1810s–1820s reached prime adult ages in the 1850s–1860s and were directly exposed to the mortality crisis associated with the Taiping civil war, mechanically depressing observed headcounts relative to the fertility-based r for these cohorts.[7]

4.2. 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 de-duplicating 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 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 (see Table 1).

Averaged over the seventeenth–nineteenth centuries, Fujian is the most male-skewed 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. The cross-province ordering follows familiar geography: stronger male surplus in the southeast (Fujian, Zhejiang, parts of Guangdong) and more balanced composition in provinces such as Jiangxi and Shandong. Taiwan's averages are consistent with a frontier setting that attracted male migrants in earlier periods, with movement toward more balanced composition in later periods (as seen in modern registers).

[7] As a timing check, the 1810 (1820) cohort is age 50 (40) in 1860.

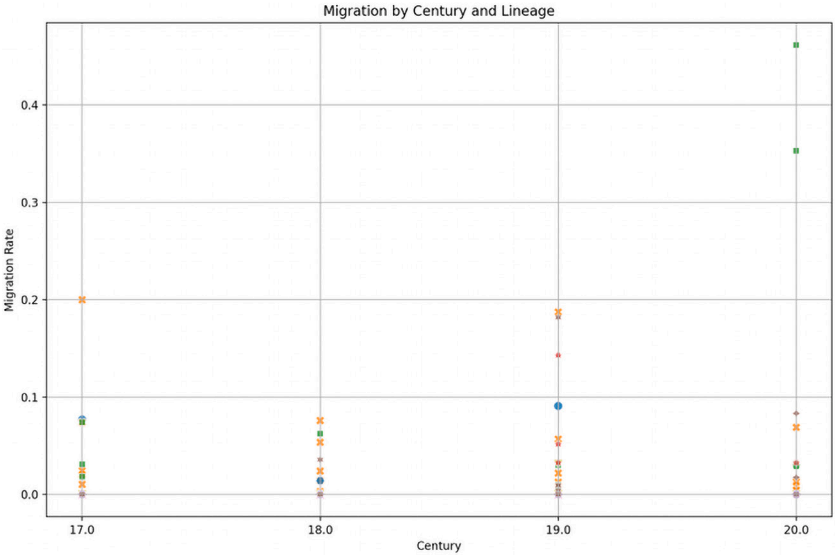

Fig. 3. Out-migration rate by century and lineage.
Note: The plot is restricted to the sixteen lineages that meet two stability criteria: (i) at least 500 geo-coded individuals in total and (ii) a minimum of ten observations in every represented generation. Each marker reports the share of persons in 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.

External anchors. Because the universe here appends wives and then de-duplicates, the male-heavy levels are not primarily due to omission of women. For independent checks, Appendix summarizes register-based anchors: Liaoning's household registers (CMGPD–LN) provide a North-China reference for adult sex composition over 1749–1909, and Taiwan's 1905 census and early-twentieth-century SRB series benchmark balanced birth ratios and a declining overall male surplus. Our Li-surname sample contains too few Liaoning observations to support a direct province-level comparison, so Liaoning is used qualitatively as a northern anchor, while quantitative contrasts rely on the provinces listed above and the Taiwan benchmarks.[8]

4.3. *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 37,239 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 37,239 geo-coded parent–child pairs, we identify 843 moves, an out-migration rate of 2.26 per cent. Fig. 3 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,[9] so minor errors do not inflate mobility measures. The resulting distribution (Fig. 4) 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.

[8] Residual biases may remain even after spouse augmentation, including survivorship of patrilines (extinct lines are absent), regional and temporal variation in recording spouses and remarriages, and cohort-specific recording intensity. These can shift levels across provinces and centuries without implying undercounting of women per se.

[9] 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.

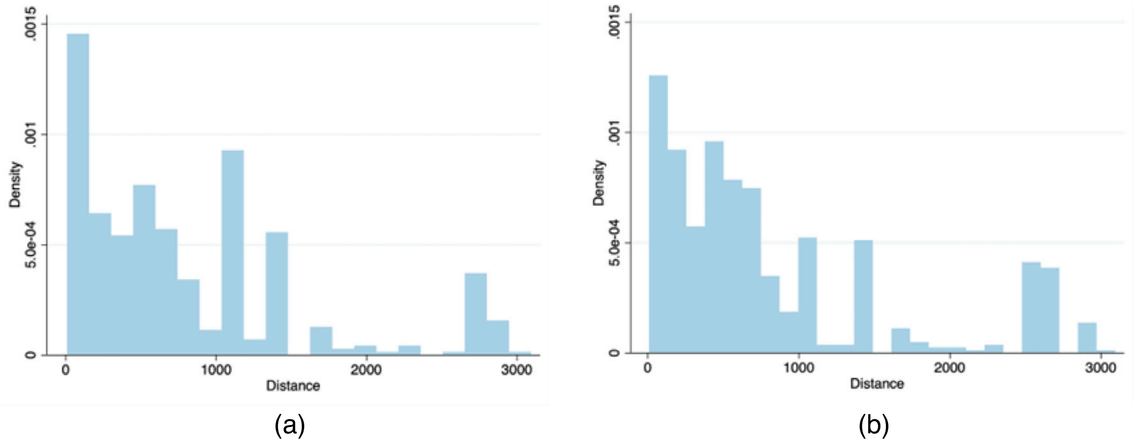

Fig. 4. Histogram of migration distances inferred from parent–child pairs.
Note: Panels plot great-circle distances (km) for all moves with measurable origin–destination pairs. The left panel uses the parent's place of death as origin; the right panel uses the parent's birthplace. Moves to overseas destinations and records with only country-level locations (e.g. "China") are excluded from these distance plots.

Table 2
Sex ratio and probability of having any recorded child.

Dependent variable: Child Dummy	(1)	(2)	(3)	(4)
	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
R^2/Pseudo R^2	0.0021	0.0015	0.016	0.011
Observations	95,613	95,613	95,613	95,613

Note: Sex ratios (males per 100 females) are computed at the province–century level from the Li-surname universe augmented with spouses after de-duplication (Section 4.2). The dependent variable equals 1 if the man has *any recorded child*; because daughters are rarely listed, this proxy effectively captures "any recorded son," so some men with only daughters may be coded as childless. Columns 1 and 3 report OLS; Columns 2 and 4 report logit. Columns 3–4 include century fixed effects. Robust standard errors clustered by province of birth in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. The sample excludes women and those born after 1900.

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.

5. Documenting expected demographic relationships

This section shows that two relationships expected in prior work are present in the genealogical panel, demonstrating that the data can support large-scale cross-regional and panel-style analyses.

5.1. Skewed sex ratios and male childlessness

We test whether male-biased sex ratios increase *male childlessness*. The outcome is a dummy for whether a man has at least one *recorded* child in the genealogy. Because daughters are seldom listed in traditional *jiapu*, this measure should be read as a proxy for having at least one *recorded son*; men with only daughters may appear childless. The sample is restricted to men.

Sex ratios are defined at the province–century level from the Li-surname universe augmented with spouses after de-duplication (Section 4.2). Each value represents the number of males per 100 females with known sex and provincial birthplace.

We regress the male-childlessness proxy (any recorded child = 0/1) on the province–century sex ratio. Columns 1–2 report OLS/logit with province-clustered robust standard errors; Columns 3–4 add century fixed effects so coefficients are identified from within-century cross-province variation. Across specifications, higher male-to-female ratios are associated with a lower likelihood that a man has any recorded child (i.e., higher childlessness). In the fixed-effects OLS (col. 3), a one-point increase in the sex ratio is associated with a 0.00171 decrease in the probability of having any recorded child. With a one–standard-deviation change in the

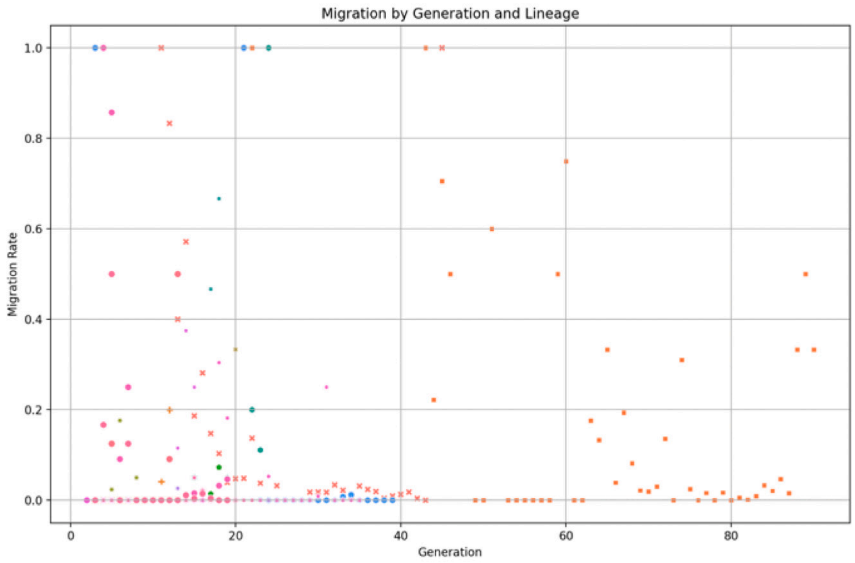

Fig. 5. Out-migration rate by generation and lineage.
Note: Each point represents the migration rate of a patrilineal generation within a specific lineage with at least 100 geocoded members. The vertical axis plots the share of male individuals in a given generation who are classified as out-migrants, 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 lineages 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.

sex ratio (≈ 16), this implies a $16 \times 0.00171 \approx 0.027$ (2.7 percentage-point) decrease. The logit estimates confirm the same pattern with comparable magnitudes.[10]

5.2. Lineage and mobility

Scholars since Freedman (1958, 1966), and more recently Watson (1988), Szonyi (2002), describe a characteristic developmental cycle for Chinese patrilineal descent groups. Founders are portrayed as mobile pioneers who leave native districts to claim land or to escape political or ecological shocks. Once a viable settlement forms, the very institutions that sustain the group — ancestral halls, graveyards, collaborative landholdings, and an active ancestral cult — tie descendants to the locality. Following the South China lineage literature, we use *lineage* to denote this localized, genealogy-based corporate group.[11] The implied pattern is high migration in the first few generations followed by progressive consolidation and spatial inertia.

We examine whether the panel recovers this predicted pattern. We code migration 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.[12] Generations are enumerated within each lineage, with generation 1 denoting the recorded ancestor at the top of the genealogy. One exceptionally large lineage spans 90 generations and includes 5,301 members; because it exerts disproportionate leverage in generation–migration regressions, we exclude it in the baseline specification (results are similar when it is included; see Appendix Table A.X).

Fig. 5 visualizes the broad pattern, restricting to lineages with at least 100 geocoded members. Founding generations display extreme heterogeneity — some lineages remain sedentary, others are entirely migrant — but the envelope of rates narrows quickly. By generation 20, virtually every lineage records fewer than five percent out-migrants; many record none. Appendix Figure A.3 traces the twelve largest lineages in our geocoded 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 3 confirms this downward trend. Regressing the binary migration indicator on generation, with century fixed effects, yields a highly significant negative coefficient (−0.0137, s.e. 0.0038). In the logit specification, this implies that each additional generation reduces the odds of out-migration by roughly 1.4 percent.

[10] We have also explored the same specification at the prefecture level using sex ratios defined for each prefecture–century cell (see Appendix Table A.1). The results are qualitatively similar to those in Table 2, but the estimates become more sensitive to outliers because many prefectures contain too few individuals to produce a reliable sex ratio. Only a subset of prefectures meets the minimum sample-size and data-quality thresholds. For this reason, we view the province level as the more appropriate unit of analysis given the current single-surname sample. Once additional surnames are incorporated, thereby enlarging local sample sizes, the full set of analyses can be replicated at the prefecture level without difficulty.

[11] 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.

[12] The coding rules follow Section 4.3. Results are virtually identical if we use paternal birth rather than death polygons.

Table 3
Effect of generation on the probability of out-migration.

VARIABLES	Dependent variable: Out-Migration (0/1)
	(1)
Generation	−0.0137*** (0.00383)
Century FE	Yes
Pseudo R^2	0.102
Observations	30 875

Note: Logit regression of out-migration on patrilineal generation, controlling for century fixed effects. Robust standard errors are shown in parentheses. One exceptionally large lineage spanning 90 generations and 5,301 members is excluded from the estimation because it exerts disproportionate leverage. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

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.

6. Conclusion and future directions

Using more than 190,000 *FamilySearch* records for a single common surname, we assemble a large, internally coherent panel suited to quantitative analysis. Descriptive series behave as expected. Male counts expand in the seventeenth and eighteenth centuries, contract in the mid nineteenth century, then partly recover. Sex ratios are strongly male-biased in places such as Fujian, elevated in parts of Henan, and near parity in Jiangxi. Migration is predominantly local with a thin long-distance tail.

To our knowledge, this is the first study to assemble genealogies with broad provincial coverage across China. The panel provides historical demography estimates for many places rather than a handful of well-documented localities. These level estimates remain subject to inherent biases in genealogical sources such as uneven preservation, under-recording of daughters, and social-class skew. We therefore present the aggregates as informative but use them with caution.

The same coverage makes the data well suited to analyses that leverage variation across places, cohorts, and generations—an approach that has become standard in modern economic history. This echoes recent work using crowd-sourced genealogies to study how outcomes vary across regions in France (Blanc, 2023). Our panel is readily linkable to commonly used regional information and can be used in familiar cross-sectional and panel specifications. The descriptive regressions here are proof of concept. Higher male-biased sex ratios are associated with a higher share of men with no recorded offspring, and deeper lineage depth is associated with lower out-migration. We focus on the direction of the estimates rather than precise magnitudes. Because the main data issues work through levels and social-class coverage, they should not bias these associations so long as distortions are unrelated to the key regressors and specifications absorb broad compositional differences with standard controls and fixed effects.

The *FamilySearch* approach complements the cover-to-cover digitization of single clan books. Full editions remain indispensable for topics such as ritual practice, property transfers, and lineage rules, but they are labor-intensive and piecemeal. The *FamilySearch* tree aggregates material from thousands of clans through a uniform interface, allowing researchers to extract macro-demographic indicators on a scale that would be prohibitively costly if digitized clan by clan.

The pipeline is readily extensible. Replicating it for further common surnames would raise the sample into the millions and enable broader analyses of Chinese historical demography. Broader surname coverage will also permit closer quantitative benchmarking against existing lineage datasets and regional demographic reconstructions. Such comparisons will make it possible to evaluate the representativeness of individual surnames and to quantify biases that remain difficult to assess with a single-patriline sample. *FamilySearch* already contains millions of Chinese records of similar structure, so expanding the analysis is straightforward. The Guangdong–Southeast Asia corridor 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 large historical microdatasets and revisit questions about family, migration, and economic change in late imperial and modern China.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eeh.2025.101734.

Data and code availability

Upon publication, we will release replication materials, including the code used for data analysis, along with a de-identified version of the dataset. All materials will be deposited in a public repository with a persistent DOI.

References

Blanc, Guillaume, 2023. The Cultural Origins of the Demographic Transition in France. Technical Report.

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

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

Calderón-Bernal, Liliana P., Alburez-Gutierrez, Diego, Zagheni, Emilio, 2023. Analysing Biases in Genealogies Using Demographic Microsimulation. Max Planck Institute for Demographic Research, MPIDR Working Paper WP-2023-034.

Campbell, Cameron D., Lee, James Z., 2002. State views and local views of population: Linking and comparing genealogies and household registers in Liaoning, 1749–1909. Hist. Comput. 14 (1–2), 9–29.

Campbell, Cameron, Lee, James Z., 2011. Kinship and the long-term persistence of inequality in Liaoning, China, 1749-2005. Chin. Sociol. Rev. 44 (1), 71–103.

Chong, Michael, Alburez-Gutierrez, Diego, Grow, André, Zagheni, Emilio, 2022. Biases in Online Genealogical Data and Methods for Their Correction: Evidence from the FamiLinx Dataset. Max Planck Institute for Demographic Research, Rostock, Germany, MPIDR Working Paper WP-2022-008.

Clark, Gregory, 2007. Farewell to Alms. Princeton University Press, Princeton, New Jersey.

Colasurdo, Agostino, Omenti, Emanuele, 2024. Using online genealogical data for demographic research: An empirical examination of the FamiLinx database. Demogr. Res. 51 (41), 1139–1172.

Dai, Ying, 2025. Lineage genealogies as a new source for researching the occupational structure of twentieth-century China: Tradition (partially) transformed. Hist. Methods: A J. Quant. Interdiscip. Hist. 58 (1), 54–79.

Fairbank Center for Chinese Studies and the Institute for Chinese Historical Geography at Fudan University, 2012. CHGIS, version: 5.

Freedman, Maurice, 1958. Lineage Organization in Southeastern China. University of London, Athlone Press, London.

Freedman, Maurice, 1966. Chinese Lineage and Society: Fukien and Kwangtung. Athlone Press, London.

Harrell, Stevan, 1987. On the holes in Chinese genealogies. Late Imp. China 8 (1), 18–35.

Harrell, Stevan, et al., 1985. The rich get children: Segmentation, stratification, and population in three Chekiang Lineages, 1550-1850. Fam. Popul. East Asian Hist. 81–109.

Hu, Sijie, Evolutionary advantage of moderate fertility during Ming–Qing China: A unified-growth perspective, J. Econom. Growth, Forthcoming).

Hu, Sijie, 2023. Descendants over 300 years: Marital fertility in five lineages in Qing China. Asia-Pac. Econ. Hist. Rev. 63 (2), 200–224.

Kaplanis, Joanna, Gordon, Assaf, Shor, Tal, Weissbrod, Omer, Geiger, Dan, Wahl, Michael, Gershovits, Maya, Markus, Boaz, Sheikh, Muhammad, Gymrek, Melissa, Bhatia, Gaurav, MacArthur, Daniel G., Price, Alkes L., Erlich, Yaniv, 2018. Quantitative analysis of population-scale family trees with millions of relatives. Science 360 (6385), 171–175.

Lee, James, Wang, Feng, Campbell, Cameron, 1994. Infant and child mortality among the Qing nobility: Implications for two types of positive check. Popul. Stud. 48 (3), 395–411.

Liu, Ts'ui-jung, 1978. The demographic structure of a Chinese lineage, 1650–1900. Popul. Stud. 32 (3), 473–489.

Price, Joseph, Buckles, Kasey, Van Leeuwen, Jacob, Riley, Isaac, 2021. Combining family history and machine learning to link historical records: The census tree data set. Explor. Econ. Hist. 80, 101391.

Shiue, Carol H., 2016. A culture of kinship: Chinese genealogies as a source for research in demographic economics. J. Demogr. Econ. 82 (4), 459–482.

Shiue, Carol H., 2017. Human capital and fertility in Chinese clans before modern growth. J. Econ. Growth 22 (4), 351–396.

Shiue, Carol H., 2025. Social mobility in the long run: A temporal analysis of Tongcheng, China, 1300 to 1900. J. Econ. Hist.

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

Standardization Administration of China, 2013. Codes for the administrative divisions of the people's republic of china (gb/t 2260). (GB/T 2260), Standardization Administration of China.

Stelter, Robert, Alburez-Gutierrez, Diego, 2022. Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics. Proc. Natl. Acad. Sci. 119 (11).

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

Telford, Ted A., 1986. Fertility and population growth in a nineteenth-century Chinese lineage. In: Lee, James Z., Campbell, Cameron, Chen, Ching-chih (Eds.), Family and Population in East Asian History. University of California Press, Berkeley, pp. 87–112.

Watson, Rubie S., 1988. Remembering the dead: Graves and politics in southeastern China. In: Watson, James L., Watson, Rubie S. (Eds.), Death Ritual in Late Imperial and Modern China. University of California Press, Berkeley, pp. 121–143.

Zhao, Zhongwei, 1994. Demographic conditions and multi-generation households in Chinese history. Results from genealogical research and microsimulation. Popul. Stud. 48 (3), 413–425.

Zhao, Zhongwei, 2001. Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. Popul. Stud. 55 (2), 181–193.