Regional Integration within National Segmentation: Multilayered Boundary Effects of Labor Market in Late Imperial China

Cheng Yang, Zeyu Chen, Yuankai Jin, Xin Fan*

Market integration is a pivotal theme in global economic history, as its role in promoting labor division, technology diffusion, and economic growth has become a key factor in addressing the Great Divergence debate (Shiue and Keller 2007; Cox 2017). Market integration in early modern Europe laid the foundation for the Industrial Revolution, whereas non-Western countries like India were hindered in their economic modernization by persistently low levels of market integration (Granger and Elliott 1967; Jacks 2004; Findlay and O'Rourke 2007; Chilosi et al. 2013; Studer 2015; Dobado-González, García-Hiernaux, and Guerrero 2015). Market integration is influenced by both formal and informal institutions, with institutional uniformity such as standardized currency systems, property laws and free trade policies playing a particularly crucial role (North and Thomas 1973; Williamson 1995; Casella 1996; Carrieri, Errunza, and Hogan 2007; Mitchener and Ohnuki 2009; Chilosi et al. 2013).

Existing research suggesting that Chinese grain market in the 18th and 19th centuries was highly integrated, approaching levels observed in Western countries (Wang 1992; Li 2000; Shiue and Keller 2007), despite China's significant lag in per capita GDP and living standards. It has been suggested that this integration is a necessary but insufficient condition for long-term development in China, and did not play a decisive role in the divergence between China and Europe. Market integration of grain prices, while significant, does not necessarily reflect the overall integration of Chinese markets, as grain was a heavily regulated commodity, and its patterns may not represent broader market dynamics. Labor, a key factor in differentiating Western and Eastern paths of economic growth (Allen 2001; Sugihara 2004), has been studied through international wage levels and labor force composition (Li and Zanden 2012; Yang 2022; Paker, Stephenson, and Wallis 2024; Liu 2024), but the integration of labor market and its effects on the Great Divergence remains underexplored.

Labor market integration is fundamental for economic development and social welfare (Wildasin 2000; Robertson 2000; Zimmermann 2009; Dorn and Zweimüller 2021). One major branch of this research focuses on boundary effects, which examine the institutional barriers to labor mobility (Helliwell 1997; Evans 2003; Gorodnichenko and Tesar 2009; Bartz and Fuchs-Schündeln 2012). Studies on modern China have identified factors such as the household registration (*hukou*) system as significant impediments to inter-regional labor migration (Bosker et al. 2012; Su, Tesfazion, and Zhao 2018). For pre-industrial China, extant scholarship presents divergent narratives regarding the constraints and facilitators of labor mobility across various regions and migration types (Entenmann 1980; Wong 1997; Pomeranz 2000; McKeown 2011), yet comprehensive data on national labor mobility are scarce, underscoring the imperative for additional empirical inquiry.

Drawing on the rarely explored but abundant homicide court records (*Xingke Tiben*), which provides individual-level migration records with high spatial and temporal resolution in 1734–1898, this study

* Cheng Yang, Institute of China's Economic Reform and Development, School of Economics, Renmin University of China; Cambridge Group for the History of Population and Social Structure, Department of Geography and Faculty of History, University of Cambridge; yangcheng8549@ruc.edu.cn. Zeyu Chen, School of Economics, Renmin University of China; rucczy@ruc.edu.cn. Yuankai Jin, Institute of Historical Geography, Peking University; 2001213349@pku.edu.cn. Xin Fan, National Research Center for Political Economy of Socialism with Chinese Characteristics, School of Economics, Renmin University of China; fanxin2020@ruc.edu.cn.

investigates labor market integration in 18th- and 19th-century China through the lens of boundary effects. By doing so, it aims to offer fresh insights into the dynamics of labor mobility and its role in historical economic development. The findings suggest a complex structure of "regional integration coexisting with national segmentation," primarily shaped by the unique multilayered administrative structure of late imperial China.

The structure of this paper is organized as follows. Section 2 introduces the historical background, theoretical model, and estimation methods, laying the groundwork for the analysis. Section 3 details the data construction and descriptive statistics, providing the empirical foundation for the study. Section 4 examines the province and *xiaqu* boundary effects on labor migration. Sections 5 and 6 explore the limited regional integrated labor market from the perspectives of real income and prefecture size distribution, highlighting the impact of administrative boundaries on income convergence and labor spatial allocation. Section 7 concludes.

HISTORICAL BACKGROUND, THEORETICAL MODEL, AND ESTIMATION

Theoretical underpinning

The principal-agent problem is a core issue in economic development. Given that information asymmetry and misaligned interests have long existed between the state and local society, as well as between rulers and the ruled (Miller 2005), nations throughout history and across the globe have consistently established administrative divisions, delegating a measure of authority to local officials (Richardson 1994; Carpenter 2005; Kunt 2014). This has led to the creation of multilayered internal administrative boundaries. These boundaries serve both positive and negative functions: they reduce internal conflicts and ensure social stability within regions, but they also hinder socioeconomic mobility and interregional connections, which results in boundary effects (Skinner 1977; Kone et al. 2018).

For countries like China, with vast territories, large populations, and a long history of centralized rule, the principal-agent problem is particularly acute. The relatively stagnant information and transportation technologies, territorial expansion, population growth, and economic development exacerbate the agent problem. Therefore, China in the 18th and 19th centuries, a period in which these processes were unfolding, presents an important case for studying the agent problem-administrative boundary dynamic and its resulting developmental and non-developmental consequences (Sng 2014; Ma and Rubin 2019). The decentralization of the administrative system and the lack of political modernization are considered key factors contributing to the Great Divergence between early modern China and the West (Wong 1997; Goldstone 2016), and research on China's internal administrative boundaries can offer valuable insights into this theoretical framework.

Evolution of Qing's political hierarchy and migration

On the history of administrative boundaries. China's multi-level administrative boundaries have a long historical tradition. The origins of China's first-level administrative system dates back to the legendary "Yugong Nine Provinces" (Mostern 2011) with its practical implementation as provincial inspector (*zhou cishi*) no later than 2000 years ago (Crespigny 2018). China's provincial administrative units were often comparable in size and population to European nations, granting their governors immense authority. When these governors commanded overwhelming military power, they became warlords, challenging centralized control and causing political fragmentation, particularly during the late Tang Dynasty, which led to the Song Dynasty's military centralization reforms (Graff 2017; Lorge 2017). Over time, the names and boundaries of

administrative divisions evolved, but the basic framework persisted, culminating in the province-prefecture-county system during Ming and Qing dynasties (Brook 1985).

As territorial expansion and population growth intensified during the Ming and Qing dynasties, the position of governors-general (*zongdu*) was established above the provincial level to oversee and manage multiple provinces. The hierarchical structure between provinces and *zongdu* gradually evolved and became institutionalized in the eighteenth and nineteenth centuries, consolidating provincial administration and mitigated fragmentation by centralizing governance under the boundaries of the governors-general, *zongdu xiaqu* (Guy 2013; Saywell and Chu 2020).

By 1760, Qing government established a stable framework of eight governor-general jurisdictions, including Zhili, Liangjiang, Minzhe, Huguang, Liangguang, Sichuan, Shangan, and Yungui.[1] The Shengjing *Jiangjun* in Northeast China also functioned similarly, consolidating administration over the three provinces. By 1906, this role transitioned into the Dong Sansheng *Zongdu*, integrating the area into the broader *xiaqu* framework.

The relationship between governor-generals and provincial governors reflected a dual structure of oversight and collaboration (Du 2009). While governor-generals maintained overarching authority over military, fiscal, and administrative matters, provincial governors handled localized civil affairs. Governor-generals and provincial governors played pivotal roles as intermediaries between the emperor and local administrations. The practice of *ti bu* (nominations for local official vacancies) granted them significant autonomy in personnel decisions (Ch'u 1962; Koss 2017). They also transmitted imperial policies to local officials and reported local conditions to the central court.

By the late nineteenth century, the Qing faced internal and external challenges that tested its centralized governance model. The rise of localism during this period reflected the growing autonomy of regional authorities (Kuhn 1980). These officials, initially tasked with suppressing rebellions and modernizing military forces, expanded their influence into economic and social reforms. They oversaw industrialization, tax reforms, and infrastructure projects, reshaping their roles beyond traditional administrative confines (Bays 1970). This shift marked the gradual erosion of central authority as local leaders became semi-autonomous power centers, a significant factor in the Qing Dynasty's eventual decline.

On migration in late imperial China. The Qing dynasty marked one of the most active periods of migration in Chinese history (Cao 1997; Rowe 2002). Existing scholarship has examined long-distance, large-scale "mainstream migration" together with multiple other types of migration on different distance levels (Cao 1997; Pomeranz 2000; Campbell and Lee 2001; McKeown 2011). Uncultivated frontiers and mountainous areas were considered the primary destinations for migrants during the Qing dynasty, contrasting with the urban-oriented migration patterns observed in contemporary Europe (Ho 1959; Gottschang and Lary 2000). However, studies also suggest that Chinese cities exerted a significant pull on migrants, many of whom engaged in non-agricultural work, ranging from skilled labor to unskilled occupations (Rowe 1993; Zelin 2006).

Formal institutions, particularly those conducted by local administrative authorities, played a role in shaping migration patterns. While encouraging some interprovincial migration through measures such as tax exemptions, the state generally adopted a laissez-faire approach to migration in most regions (Entenmann 1980). Additionally, certain administrative areas were designated as restricted zones where migration was limited, although these restrictions gradually weakened over time (Schlesinger 2021; Wang 2021).

Informal institutional factors also significantly influenced migration and the choice of destinations. Among

[1] (Qianlong) "Da Qing Hui Dian" Volume 4, Ministry of Personnel, Official System 4, "*Zongdu* is of the second rank, with Shangshu title of the first rank, one Zhili, one Jiangnan and Jiangxi, one Fujian and Zhejiang, one Hubei and Hunan, one Shaanxi and Gansu, one Guangdong and Guangxi, one Sichuan, and one Yunnan and Guizhou."

these, kinship organizations played a particularly prominent role. As powerful actors in local society, clans often acted as networks providing financial support, technical training, and employment opportunities for migrants (Gillette 2016; de la Croix, Doepke, and Mokyr 2018). Furthermore, these organizations offered crucial protections for migrants seeking to settle in their new destinations, helping to ensure their security and fortune, even at times at the cost of ethnic conflicts (Shepherd 1993; Leong and Skinner 1997).

One of the primary challenges in migration studies lies in the limitations of available data sources. Existing research often relies on either national-level qualitative data such as anecdotes and archives, or localized quantitative records such as household registrations and lineage genealogies (Campbell and Lee 2001). However, there remains a lack of nationwide individual-level quantitative data that would allow for comparative analyses across different regions and different types of migrants, as well as assessments of institutional influences on the scale and scope of migration.

Theoretical model

To guide our empirical analysis, we develop a static discrete choice model in which individuals select locations based on income, prices, and migration costs. The model builds upon existing work on spatial sorting (for example, Ahlfeldt et al. 2015; Tombe and Zhu 2019), with a specific emphasis on the effects of multilayered administrative boundaries. We then derive a structural gravity equation from this model, which allows for estimating the administrative boundary effects.

Consider a closed economy consisting of L individuals, each supplying a single unit of labor. The economy comprises N distinct locations, denoted by $o, d \in \mathcal{N} \equiv \{1, \cdots, N\}$, where the subscript od represents migration from o to d.

Individuals determine migration decisions by maximizing their utility. Suppose that they are risk neutral and the utility function of individual i migrating from o to d is

$$U_{od}^{i} = \frac{\epsilon_{od}^{i}}{\mu_{od}} C_{od}^{i}, \tag{1}$$

where ϵ_{od}^{i} represents an idiosyncratic component of utility that is assumed to be independent of both migration costs and the characteristics of the locations. The term C_{od}^{i} is a consumption index, which is modeled as a constant elasticity of substitution (CES) aggregation over the set of differentiated commodities Ω, characterized by an elasticity of substitution $\eta > 1$:

$$C_{od}^{i} = \left[\sum\nolimits_{\omega \in \Omega} c_{od}^{i}(\omega)^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}. \tag{2}$$

The term μ_{od} represents the bilateral resistance to migration from o to d, which encompasses not only the physical and psychological of living far from one's original location but also the utility loss arising from institutional restrictions. Given our primary focus on estimating administrative boundary effects, it is crucial to isolate some predetermined resistance factors that are strongly correlated with crossing such boundaries, particularly distance-related and culture-related costs. Specifically, μ_{od} is modeled as the product of five distinct components:

$$\mu_{od} = \bar{d}_{od} \tilde{d}_{od} b_{od}^{p} b_{od}^{x} \lambda_{od} \geq 1, \tag{3}$$

where $\bar{d}_{od}$, $\tilde{d}_{od}$, b_{od}^{p}, and b_{od}^{x} represent distance-related migration costs, culture-related migration costs, province boundary effects, and *xiaqu* boundary effects, respectively. Additionally, λ_{od} captures the unobserved disutility, which is assumed to be uncorrelated with other resistance factors.

Define the nominal income that individual i could earn in location d as I_d^i, expressed as $I_d^i = z_d^i \bar{I}_d$.[2] Here, $\bar{I}_d$ represents the average income across all industries and occupations, encompassing various resources such as wages, self-employment income, and transfer payments. The idiosyncratic term z_d^i captures the deviation of individual income in location d from this average due to different jobs and skills. With prices of commodities denoted as $p_d(\omega)$ for commodity ω, the budget constraint of individual i migrating from o to d is

$$\sum\nolimits_{\omega\in\Omega} c_{od}^i(\omega) p_d(\omega) \le z_d^i \bar{I}_d. \tag{4}$$

Using the properties of CES aggregation, it can be shown that the indirect utility function takes the following form:

$$V_{od}^i = \frac{\nu_{od}^i \bar{I}_d}{\bar{d}_{od}\tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od} P_d}, \tag{5}$$

where $\nu_{od}^i \equiv \epsilon_{od}^i z_d^i$ is a combined idiosyncratic term of individual i's preference for location d, and P_d denotes the aggregated price index in location d (see Appendix C for a formal derivation).

The migration decision for individual i from location o involves selecting the destination d that maximizes the utility. Since an individual's idiosyncratic preferences are observable only to the individual but unobservable to researchers, they are typically assumed to be drawn from a joint distribution $\mathcal{F}$. Consequently, the probability of individual i choosing location d is given by

$$\begin{aligned} m_{od}^i &= \Pr\left\{V_{od}^i \ge \max_{d'\neq d}\{V_{od'}^i\}\right\} \\ &= \Pr\left\{\frac{\nu_{od}^i \bar{I}_d}{\bar{d}_{od}\tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od} P_d} \ge \max_{d'\neq d}\left\{\frac{\nu_{od'}^i \bar{I}_{d'}}{\bar{d}_{od'}\tilde{d}_{od'} b_{od'}^p b_{od'}^x \lambda_{od'} P_{d'}}\right\}\right\}, \end{aligned} \tag{6}$$

where $\{\nu_{od}^i\}_{d\in\mathcal{N}} \sim \mathcal{F}$.

For tractability, we follow the standard approach in the spatial sorting literature (Redding and Rossi-Hansberg 2017) and assume that the idiosyncratic term ν_{od}^i is independently and identically drawn from a Fréchet distribution, $F_\nu(x) = \mathrm{e}^{-x^{-\kappa}}$, with $x > 0$, where the shape parameter $\kappa > 1$ governs the degree of dispersion across individuals. Under this assumption, the closed-form expression for the choice probability can be derived as follows (see Appendix C for the formal derivations of the following two probabilities):

$$m_{od} = \frac{\bar{I}_d^\kappa P_d^{-\kappa}\left(\bar{d}_{od}\tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od}\right)^{-\kappa}}{\sum_{d'\in\mathcal{N}} \bar{I}_{d'}^\kappa P_{d'}^{-\kappa}\left(\bar{d}_{od'}\tilde{d}_{od'} b_{od'}^p b_{od'}^x \lambda_{od'}\right)^{-\kappa}}. \tag{7}$$

Since each individual's idiosyncratic term is independently drawn from the same distribution, the law of large numbers implies that the proportion of individuals who migrant to location d (or stay locally if $d = o$) among all individuals originating from location o will converge to this probability.

To match our migrant data, we derive the conditional probability that individual i from location o, given the decision not to remain locally, migrants to location d as follows:

$$\begin{aligned} m_{od,-o} &= \Pr\left\{V_{od}^i \ge \max_{d'\neq d}\{V_{od'}^i\} \,|\, V_{oo}^i < \max_{d'\neq o}\{V_{od'}^i\}\right\} \\ &= \frac{\bar{I}_d^\kappa P_d^{-\kappa}\left(\bar{d}_{od}\tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od}\right)^{-\kappa}}{\sum_{d'\neq o} \bar{I}_{d'}^\kappa P_{d'}^{-\kappa}\left(\bar{d}_{od'}\tilde{d}_{od'} b_{od'}^p b_{od'}^x \lambda_{od'}\right)^{-\kappa}}. \end{aligned} \tag{8}$$

Similarly, given a sufficient number of migrants, the proportion of individuals who migrant to location d among all migrants originating from location o will converge to $m_{od,-o}$.

[2] In standard quantitative spatial models, income and the spatial allocation of labor are typically determined simultaneously in equilibrium. However, since our model does not explicitly incorporate the production side, income is treated as exogeneous when solving for the optimal migration decision. In this context, income should be understood as the equilibrium income.

Estimation

We interpret the observed choices of migration destinations as a finite sample from the data generating process outlined in the above discrete choice model. Accordingly, Equation (8) bridges the model with our migrant data. To identify administrative boundary effects, we first need to explicitly model each resistance factor.

Firstly, crossing administrative boundaries is typically correlated with longer migration distances. Consequently, neglecting distance-related costs could lead to significant overestimations of boundary effects. The distance-related term $\bar{d}_{od}$ captures at least two distinct types of costs: longer migration distances generally incur higher transportation expenses and increased information incompleteness.[3] Following the common approach in the literature (Tombe and Zhu 2019), we assume that the cost associated with physical distance follows a constant elasticity; that is,

$$\bar{d}_{od}^{-\kappa} = \bar{\mu} \cdot Distance_{od}^{\sigma}, \tag{9}$$

where parameter $\bar{\mu}$ controls for the measurement units and σ is the distance elasticity of migration.[4]

Additionally, the culture-related term $\tilde{d}_{od}$ captures the security and support derived from familial ties, consistent with prior studies highlighting the crucial role of kin-based organizations in pre-modern China (Kumar and Matsusaka 2009; Greif and Tabellini 2017). Let $Culture_{od}$ denote the similarity in clan composition between locations o and d, which is subsequently quantified using the cosine similarity of surname distributions, taking values between zero and one. To ensure that the associated migration costs are no less than one, they are modeled using the following exponential function:

$$\tilde{d}_{od}^{-\kappa} = \tilde{\mu} \cdot \mathrm{e}^{\pi(1-Culture_{od})}, \tag{10}$$

where parameter π controls the size of culture-related costs.

Finally, b_{od}^{p} and b_{od}^{x} represent the costs associated with crossing administrative boundaries. Notably, since *xiaqu* is a higher administrative hierarchy that typically encompasses multiple provinces, crossing a *xiaqu* boundary inherently involves crossing provincial boundaries. Therefore, b_{od}^{x} does not represent the total cost of crossing a *xiaqu* boundary, but rather the additional cost incurred after accounting for the costs already captured by b_{od}^{p}. Formally, we define

$$\begin{aligned} \left(b_{od}^{p}\right)^{-\kappa} &= \left(\bar{b}^{p}\right)^{\mathbb{I}\{Prov_o \neq Prov_d\}}, \\ \left(b_{od}^{x}\right)^{-\kappa} &= \left(\bar{b}^{x}\right)^{\mathbb{I}\{Xiaqu_o \neq Xiaqu_d\}}, \end{aligned} \tag{11}$$

where $\mathbb{I}\{\cdot\}$ takes the value of 1 if the condition within the braces holds. Here, the costs associated with crossing different *xiaqu* boundaries are initially assumed to be constant ($\bar{b}^{x}$), as are the costs of crossing provincial boundaries. However, we will allow these costs to vary with respect to different destinations in the subsequent analysis.

Substituting Equations (9) to (11) into Equation (8) gives the following equation:

$$\begin{aligned} m_{od,-o} = \exp[\, &\gamma + \alpha_o + \beta_d + \sigma \ln Distance_{od} + \pi(1 - Culture_{od}) \\ &+ \left(\ln \bar{b}^{p}\right)\mathbb{I}\{Prov_o \neq Prov_d\} + \left(\ln \bar{b}^{x}\right)\mathbb{I}\{Xiaqu_o \neq Xiaqu_d\}] + \varepsilon_{od}, \end{aligned} \tag{12}$$

[3] If we narrow the information incompleteness to income information, the standard approach is to assume that individuals are risk-averse and maximize their expected utility. In Appendix C, we demonstrate that, in some cases, the influence of imperfect income information can be equivalently treated as a discount factor on utility. For simplicity, we combine this disutility with the transportation expenses in Equation (9).

[4] It is important to note that some studies in urban economics model distance-related costs as an exponential function of commuting time or physical distance, which implies that the marginal cost of distance increases rapidly (for example, Ahlfeldt et al. 2015; Anagol et al. 2021; Tsivanidis 2023). However, this assumption of a rapid increase in marginal costs may not be realistic for cross-regional migration. Given that the specific modeling of distance-related costs can significantly influence estimates of boundary effects, we test both specifications in the empirical section.

where γ is a constant, origin fixed effects $\alpha_o \equiv \ln\left[\sum_{d' \neq o} \bar{I}_{d'}^{\kappa} P_{d'}^{-\kappa} \left(\bar{d}_{od'} \tilde{d}_{od'} b_{od'}^{p} b_{od'}^{x} \lambda_{od'}\right)^{-\kappa}\right]$, and destination fixed effects $\beta_d \equiv \kappa \ln(\bar{I}_d / P_d)$. The error term is $\varepsilon_{od} \equiv \mathrm{e}^{l}(\lambda_{od}^{-\kappa} - 1)$, where l represents the polynomial within the exponential function of Equation (12). Notably, Equation (12) provides a generalized gravity equation that links migrant shares to origin-specific characteristics (α_o), destination-specific characteristics (β_d), and a set of bilateral resistance factors. Of particular interest is the estimation of $\ln \bar{b}^{p}$ and $\ln \bar{b}^{x}$, which quantify the resistance imposed by administrative boundaries.

Two methods are commonly employed to estimate multiplicative equations like Equation (12). The first involves taking the logarithm of both sides of the equation and restricting the sample to those where the migrant share $m_{od,-o}$ is strictly positive. Under the assumption $\mathbb{E}[\ln \lambda_{od}^{-\kappa} | X_{od}] = 0$, where X_{od} represents all included regressors, this log-linear equation can be estimated using Ordinary Least Squares (OLS) to produce unbiased estimates. The second method uses the Poison Pseudo Maximum Likelihood (PPML) estimator, based on the assumption $\mathbb{E}[\lambda_{od}^{-\kappa} | X_{od}] = 1$, which implies $\mathbb{E}[\varepsilon_{od} | X_{od}] = 0$. As noted by Silva and Tenreyro (2006), these two assumptions are not equivalent, meaning that OLS and PPML often yield significantly different parameter estimates.

In this study, we adopt the second assumption, which naturally arises from the premise that unobserved disutility is uncorrelated with other resistance factors, and employ the PPML estimator (for similar applications, see Barjamovic et al. 2019; Caliendo et al. 2021; Severen 2023). This approach also addresses the substantial loss of observations that arises from excluding zero migrant share samples, a limitation inherent in log-linearization.[5]

DATA CONSTRUCTION AND DESCRIPTIVE STATISTICS

Accurately estimating multilayered administrative boundary effects requires data on migrants across administrative units with sufficient spatial granularity, posing a significant challenge to our study. To address this, we leverage three primary data sources: individual-level data extracting from *XKTB* reports, administrative data provided by Chinese Historical Geographic Information System (CHGIS), and prefectural population data estimated by Cao (2001). In this section, we provide a overview of the datasets, variable construction, and a discussion of potential biases, with detailed contents in Appendix B.

The *XKTB* reports are homicide trial records from the Qing Dynasty, documenting extensive demographic and economic information not only about the accuser, the accused, and accomplices, but also about witnesses and others involved.[6] To the best of our knowledge, this is one of the first datasets from the Qing Dynasty that provides individuals' residency and registration information with granularity down to the county level, while covering most of the prefectures. Although not census data, homicide records in other countries have also been utilized to investigate migration patterns, particularly in contexts of limited historical data availability (Clark 1979; Bailey 2023).

However, unlike standardized statistical data, the structure of the *XKTB* reports is much less uniform,

[5] In practice, we use the Stata package "ppmlhdfe," developed by Correia et al. (2020). In Appendix D, we compare the OLS and PPML estimators, provide the derivation of the PPML estimator, and discuss the relationship between the PPML estimator and our theoretical model. As highlighted by Sotelo (2019), PPML serves as a convenient tool for estimating the structural gravity equation without introducing additional assumptions. We also consider concerns raised by Dingel and Tintelnot (2020), who note that estimation biases may arise when the sample size of individual data is smaller than the number of location pairs. To address this, we conduct Monte Carlo simulations based on the data-generating process outlined in our theoretical model, constructing simulated datasets that mirror the size of our observed data. We find that the PPML estimator yields parameter estimates that closely align with our predetermined values.

[6] Additional historical background on the *XKTB* reports is provided in Appendix A.

requiring considerable effort to manually review each report to extract relevant information. For this study, our team collected approximately 90% of the *XKTB* reports from three periods—1761–1770, 1821–1830, and 1881–1890—resulting in a total of 52,756 reports. After data cleaning, 46,169 reports containing usable information remained for analysis. From these, we identified 126,366 individuals, including 115,773 local residents and 10,593 migrants, and construct two datasets for estimations—the prefecture-pair dataset without local pairs and the province-pair dataset with local pairs.[7]

Although the *XKTB* dataset offers extensive individual-level samples with robust spatial granularity and coverage, its exclusive focus on homicide cases may introduce biases. We address concerns about its representativeness in detail in Appendix B.3 and provide a summary here for brevity. To mitigate biases from varying homicide rates across prefectures, we reweight the sample using estimated prefectural populations from Cao (2001).

Moreover, we address sample selection biases from four aspects. First, regarding wealth and class, while individual income data is unavailable, individuals involved in homicides—whether as murderers, victims, or bystanders—exhibit only modest differences in occupational structures, suggesting relatively small wealth or class bias. Second, for age, the *XKTB* dataset shows a higher mean age and more concentrated distribution compared to population estimates from other studies. However, as this group plays a dominant role in labor supply, the selection may align well with our estimation. Third, in terms of gender, about 90% of individuals recorded in the *XKTB* reports are male. Although female samples are excluded due to data limitations, a comparison of migration distances reveals similar distributions between males and females, indicating minimal gender bias on estimations. Finally, we examine regional variations in the composition of murderers, victims, and others involved in homicide cases and find these differences to be relatively small.

Furthermore, we address concerns regarding the potential deterioration in the quality of homicide records over time, offering three possible explanations for this trend. While national-level shocks may have influenced documentation in the *XKTB* reports, our boundary effects estimation relies on comparisons among location pairs. As such, these shocks are unlikely to systematically bias our results.

Finally, we validate the *XKTB* dataset by cross-comparing with independent micro and macro sources on several derived measures, such as in-migration rates, in-migrant origin distributions in Taiwan, and out-migrant destination distributions from Hengyang Prefecture. These comparisons reveal a strong consistency, and any potential bias is likely minimal, further reinforcing the reliability of the dataset.

Table 1 provides summary statistics. Among all prefecture pairs, the mean number of migrants is 0.04. Similar to modern migration data at fine spatial scales, many pairs do not report any migrants: of the 100,172 potential migration directions, only 1,525 show a positive migrant flow in the first sample period. The average migrant flow for these pairs is 2.64, with a large dispersion (standard deviation of 5.33). For province pairs, we also included local pairs to ensure greater variation for identification of boundary effects (discussed later). This coarser spatial scale results in a much larger proportion of pairs with positive migrant flows.

We calculate the observed migrant shares (m_{od}^{Data} or $m_{od,-o}^{\text{Data}}$) based on counts of individuals recorded in the *XKTB*. However, due to significant variation in the ratio of recorded individuals to the actual population across prefectures—partly influenced by differences in homicide rates—we apply reweighting to each sample. Specifically, we calculate how many individuals each recorded in the *XKTB* represents for each prefecture, based on population estimates from Cao (2001). These ratios are then used as weights to calculate the migrant

[7] In the Qing Dynasty, household registration, based on the *lijia* (similar to living quarters in a village), was the primary unit for population management, recording each individual's birthplace. According to the *Daqing Huidian–Hubu* (the Qing Empire's regulatory code for officials, Ministry of Revenue), migrants often encountered significant delays in obtaining local household registration in their new locations—officially set at 20 years but sometimes extending much longer. This delay creates discrepancies between registration and residency, allowing us to identify migrants.

shares.

Additionally, to measure the clan culture differences between each location pair, we extract individuals' surnames and construct a vector representing the frequency of each surname for each prefecture and province. The cosine similarity between these surname vectors is then calculated as a measure of $Culture_{od}$.

The second data source used in our study is the CHGIS, which provides information to identify the administrative boundaries and central points of each prefecture and province. It also allows us to delineate the range of each *xiaqu* based on the provinces they encompass. According to Zhou et al. (2013), by the late eighteenth and nineteenth centuries, provincial boundaries and *xiaqu* divisions remained largely stable, except for the establishment of Taiwan Province from Fujian in 1885. Only minor adjustments to prefectural boundaries occurred in a few frontier provinces. Consequently, we uniformly apply the 1820 spatial data layers from the CHGIS Version 6 Dataverse across all sample periods.[8] Additionally, we identify the range of each *xiaqu* based on its encompassing of provinces.

Finally, we utilize the prefectural population data estimated by Cao (2001), a dataset widely employed in studies examining population distribution and long-term evolution in China (Chen and Kung 2016; Broadberry, Guan, and Li 2018). This dataset serves two purposes in our analysis: reweighting our samples, as previously discussed, and examining the spatial allocation of labor to evaluate the potential influence of multilayered administrative boundary effects.

Table 1 Descriptive statistics

	1761–1770			1821–1830			1881–1890		
	Obs.	Mean	S. D.	Obs.	Mean	S. D.	Obs.	Mean	S. D.
Panel A. XKTB reports									
# of *XKTB* reports	15,135			18,593			12,441		
Total individuals	45,529			50,774			30,063		
Total migrants	4,559			3,863			2,171		
Panel B. Prefecture pairs									
# of migrants	100,172	0.040	0.733	100,172	0.034	0.504	100,172	0.018	0.298
# of migrants ($l_{od} > 0$)	1,525	2.644	5.329	1,421	2.391	3.501	878	2.096	2.406
Migration distance (km)	100,172	1,358	836.8	100,172	1,358	836.8	100,172	1,358	836.8
Clan culture difference	75,900	0.601	0.214	79,806	0.573	0.213	76,452	0.669	0.214
Panel C. Province pairs									
# of individuals	529	85.15	466.0	529	95.00	537.3	529	56.22	370.5
# of migrants ($o \neq d$)	506	5.101	17.29	506	4.583	15.48	506	2.563	9.54
# of migrants ($l_{od} > 0$)	229	11.27	24.34	215	10.79	22.32	163	7.96	15.50
Migration distance (km)	529	1,368	935.4	529	1,368	935.4	529	1,368	935.4
Clan culture difference	529	0.336	0.230	529	0.314	0.241	529	0.312	0.211

Notes: This table provides descriptive statistics for each sample period. The prefecture-pair dataset excludes local pairs, while the province-pair dataset includes them. In Panel B, the second row is restricted to pairs with positive migrant flows, similar to the third row in Panel C. The first row of Panel C includes all pairs while the second row focuses on non-local pairs.

PROVINCE AND *XIAQU* BOUNDARY EFFECTS

Estimating Boundary Effects

We begin by estimating Equation (12) using the prefecture-pair dataset and the PPML estimator, comparing intra-province migration, inter-province and intra-*xiaqu* migration, and inter-*xiaqu* migration to identify multilayered boundary effects.[9] Table 2 presents the estimation results with standard errors two-way

[8] *CHGIS, Version: 6*. (c) Fairbank Center for Chinese Studies of Harvard University and the Center for Historical Geographical Studies at Fudan University, 2016.

[9] In Appendix Table F.2, we compare the estimates obtained using the PPML estimator and the OLS estimator. When applying log-linearization and OLS, the effective number of prefecture pairs in the estimation drastically drops to 3,564, representing only 2.14% of the pairs included in the PPML estimation. Although the OLS-based estimates indicate boundary

clustering at the origin and destination prefectures.[10] With migrant data from three periods—1760s, 1820s, and 1880s—we first estimate the equation with three cross-sectional datasets. Columns (1) to (3) adhere to the specification in Equation (12), including two dummies indicating whether a migration crosses a *xiaqu* boundary or a province boundary, along with controls for the logarithm of distance, the cultural difference, and origin and destination fixed effects. Columns (4) and (5) combine the three cross-sectional datasets into a panel. Column (4) introduces period fixed effects to control for macroeconomic shocks across periods, with assuming that the marginal effects of physical distance, clan differences, and the overall influence of location-specific characteristics (captured by origin and destination fixed effects) remain constant over time. In Column (5), we adopt a flexible specification to relax this assumption by interacting all controls and fixed effects with period fixed effects. As a result, the boundary effects estimated in Column (5) can be regraded as a weighted average of those in Columns (1) to (3).

Table 2 The province and *xiaqu* boundary effects

	Dependent variable: Migrant share among all out-migrants				
	1760s	1820s	1880s	All three periods	
	(1)	(2)	(3)	(4)	(5)
Crossing *xiaqu* boundary	-0.396**	-0.398**	-0.312	-0.326***	-0.372***
	(0.173)	(0.168)	(0.246)	(0.114)	(0.123)
Crossing province boundary	0.277	0.298*	0.258	0.257**	0.279**
	(0.175)	(0.169)	(0.237)	(0.104)	(0.116)
Logarithm of distance	-2.205***	-2.520***	-2.404***	-2.340***	
	(0.101)	(0.095)	(0.112)	(0.059)	
Clan difference	-1.468**	-1.083**	-0.424	-0.434***	
	(0.572)	(0.548)	(0.558)	(0.166)	
Constant	9.269***	10.880***	10.061***	9.405***	10.077***
	(0.527)	(0.494)	(0.636)	(0.319)	(0.354)
Origin fixed effects	Yes	Yes	Yes	Yes	No
Destination fixed effects	Yes	Yes	Yes	Yes	No
Period fixed effects	No	No	No	Yes	No
Origin FEs × Period FEs	No	No	No	No	Yes
Destination FEs × Period FEs	No	No	No	No	Yes
Controls × Period FEs	No	No	No	No	Yes
# of migrants in *XKTB*	4,027	3,393	1,839	9,259	9,259
# of prefecture-pair observations	58,566	62,224	45,539	186,682	166,329

Notes: This table provides estimation results using the PPML estimator. Robust standard errors two-way clustering at the origin and destination prefectures, are shown in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

All five columns present similar magnitudes for the *xiaqu* boundary effects, although the coefficients for the third period are statistically less significant. The latter is likely attributable to the smaller sample size of migrants in *XKTB* during the third period, which reduces the sample's ability to dilute noise arising from idiosyncratic preferences, consequently leading to higher standard errors. Additionally, we observe a smaller point estimate for the third period, which may be partially explained by the population mobility and the weakening of administrative control resulting from the political disruptions caused by the Taiping Rebellion (1851–1864).

Holding other factors constant, the relative change in migrant share due to crossing a *xiaqu* boundary is

$$\frac{\mathbb{E}\left[m_{od,-o}|\tilde{X}_{od}, \mathbb{I}\{Xiaqu_o \neq Xiaqu_d\} = 1\right]}{\mathbb{E}\left[m_{od,-o}|\tilde{X}_{od}, \mathbb{I}\{Xiaqu_o \neq Xiaqu_d\} = 0\right]} - 1 = \mathrm{e}^{\ln \bar{b}^x} - 1,$$

effects in the same direction as those estimated with the PPML approach, the point estimates are significantly smaller and lack statistical significance.

[10] We report two-way clustering at the origin and destination provinces in Appendix Table F.3, which accounts for more correlations among error terms but risks having too few clusters (about 20 provinces). The province-level standard errors are only slightly larger than the prefecture-level ones, so we report two-way clustering at the prefecture level in the main text for better efficiency.

where $\tilde{X}_{od}$ represents all regressors excluding $\mathbb{I}\{Xiaqu_o \neq Xiaqu_d\}$. Since Column (5) provides an estimate of -0.372 for $\ln \bar{b}^x$, the corresponding relative change is $e^{-0.372} - 1 = -31.06\%$, indicating a significant effect of *xiaqu* boundaries in deterring migration. The structural migration choice model also links this estimate to a specific parameter. Given $\mathbb{I}\{Xiaqu_o \neq Xiaqu_d\} = 1$, we have $(b_{od}^x)^{-\kappa} = \bar{b}^x$, so the utility discount factor is $b_{od}^x = e^{-\ln \bar{b}^x/\kappa}$. To recap, κ is the scale parameter of the Fréchet distribution from which individuals' idiosyncratic preferences are drawn. Here, we initially adopt the value of 2.5, with further estimation and discussion presented in the section on real income. Using this value, crossing a *xiaqu* boundary corresponds to an average utility discount of $e^{-(-0.372)/2.5} = 1.16$.

It is noteworthy that province boundaries promote migration, as this is contrary to the frequent reports of significant province boundary effects in contemporary China (for example, Su et al. 2018; Wang et al. 2023). Based on the estimates in Column (5), within a *xiaqu*, there are, on average, 32.18% more migrants choosing to migrate across provinces compared to those migrating to other prefectures within the same province.[11] However, these estimations rely on a smaller number of pairs representing inter-provincial and intra-*xiaqu* migration. Later, we will demonstrate that the "promoting effect" primarily results from a few potential outlier pairs, and excluding them renders the province boundary effects statistically insignificant. Nevertheless, our estimates highlight the distinctive boundary effects across different administrative layers: *xiaqu* boundaries deter migration, while province boundaries appear to have a softer or even promoting effect.

Additionally, we observe consistent patterns regarding the influence of migration distance and clan cultural differences across all these estimates. Firstly, longer physical distances are found to deter migration. Specifically, our analysis yields a larger distance elasticity compared to the range of -1.51 to -1.39 estimated by Tombe and Zhu (2019) for contemporary China. This discrepancy can be naturally attributed to the higher transportation costs and more pronounced information incompleteness associated with the relatively underdeveloped transportation and information infrastructure of the Qing Dynasty. Additionally, our estimates for clan culture are consistent with an emerging empirical literature on clan networks and migration, which suggests that clan networks facilitate migration (Foltz, Guo, and Yao 2020). In line with this perspective, our analysis reveals that migration is less frequent between prefectures with dissimilar clan compositions.

Robustness Check

To check the reliability of the baseline estimates, particularly regarding the "promoting effects" of province boundaries, we start by conducting four robustness tests, focusing on a specific group of migrants, our reweighting of migrant counts, the selection of the dependent variable, and the specification of migration distance.

Firstly, we observe an increase in migrants undertaking long-distance journeys, particularly those exceeding 1,100 kilometers, as shown in Appendix Figure F.1. This pattern is consistent with existing studies highlighting long-distance, centrifugal, and rural-oriented migration in eighteenth- and nineteenth-century China (Gottschang and Lary 2000; Rowe 2002). These migrations often targeted mountainous regions or remote frontiers along the empire's periphery, partly driven by specific policies promoting migration and subsequent agricultural reclamation (Rowe 2002). As a result, most of these samples represent *zongdu*-crossing migration but do not necessarily reflect the typical deterrent effects of *zongdu* boundaries, potentially influencing the estimates of *zongdu* boundary effects. Additionally, they affect the fit of the distance and, consequently, the estimates of province boundary effects. In Appendix Table F.4, we exclude prefecture pairs

[11] Specifically, $e^{0.279} - 1 = 32.18\%$.

with migration distances exceeding 1,100 kilometers and recalculate the migrant share, where the results in Column (1) presents consistent estimates.[12]

Secondly, we verify that our estimated results are not driven by our reweighting of migrant counts based on the ratio of Cao's (2001) population estimates to *XTKB* individual numbers. In Column (1) of Appendix Table F.5, we re-estimate Equation (2) using the migrant share without reweighting and find consistent results.

Thirdly, while our choice of migrant share as the dependent variable is rooted in the theoretical model, we also estimate the model using the number of migrants as the dependent variable, following some prior studies (Guo et al. 2024; Peeters 2012). As shown in Column (1) of Appendix Table F.6, our findings again remain robust.

Finally, we address concerns about the potential misspecification of migration distance, a key confounding factor in identifying boundary effects. Specifically, boundary-crossing migration is often associated with longer distances, which increase costs and deter movement. If distance is not adequately controlled, its influence remains in the error term and biases the estimation. To address this, we approximate the true distance specification using a family of polynomial functions ranging from order 2 to 6. As shown in Appendix Table F.7, the results from higher-order polynomials are close to those obtained using the logarithmic transformation of distance.

After addressing the above concerns, we turn to another challenge that has garnered increasing attention in recent years—the sample size of individual data. Dingel and Tintelnot (2020) highlight that when the sample size is small relative to the number of location pairs—especially in studies with fine spatial scales—idiosyncratic preferences may not average out, potentially leading to biased estimates. For instance, if a small number of migrants from specific origin prefectures are concentrated in a single or a few homicide cases in another prefecture of a different province within the same *xiaqu*, this can create cross-province, intra-*xiaqu* prefecture pairs with a migrant share of one, thereby potentially producing the "promoting effects."

Accordingly, a preliminary approach is to exclude potential "outlier pairs" where the origin prefecture has an inter-province pair with a migrant share of one. We filter out these pairs and detail them in Appendix Table F.8, categorizing them into two types: (1) origin prefectures with a inter-province, intra-*xiaqu* pair exhibiting a migrant share of one—these are the most likely candidates for "promoting effects"; and (2) origin prefectures with inter-*xiaqu* pairs showing a migrant share of one. We excluded these potential outliers and re-estimate Equation (12) in Table 3, where Column (1) excludes the first type and Column (2) excludes both. While we still observe a positive coefficient for province boundary effects, the point estimates decrease sharply and lose statistical significance.[13]

Table 3 Robustness check with respect to potential outlier prefecture pairs

	Dependent variable: Migrant share among all out-migrants	
	Excl. inter-province, intra-*xiaqu* outliers	Excl. inter-province outliers
	(1)	(2)
Crossing *xiaqu* boundary	-0.219*	-0.335***
	(0.115)	(0.111)
Crossing province boundary	0.108	0.065
	(0.116)	(0.110)
Origin FEs × Period FEs	Yes	Yes

[12] Regarding the structural model, this practice corresponds to the probability conditional on destinations within 1,100 kilometers. Following a similar approach to the derivation of Equation (8), the properties of the Fréchet distribution enable us to derive an analogous expression for the conditional probability, requiring only a modification of the denominator to the summation of all available destinations within 1,100 kilometers. For estimation, the remaining task is to reconstruct the migrant share and perform the estimation using the same procedure.

[13] We also conduct the aforementioned robustness checks using samples that exclude potential outlier pairs, as detailed in Appendix Tables F.4, F.5, and F.6. All these practices yield positive point estimates of province boundary effects, but most become less pronounced, with some approaching zero and lacking statistical significance.

Destination FEs × Period FEs	Yes	Yes
Controls × Period FEs	Yes	Yes
# of migrants in *XKTB*	9,237	9,177
# of prefecture-pair observations	162,705	153,121

Notes: This table provides estimation results using the PPML estimator. All regressions include constant terms that are not listed in the table. Robust standard errors, with two-way clustering at the origin and destination prefectures, are shown in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

As noted in Appendix Table F.8, all the potential outlier pairs identified above have a small number of migrants recorded in the *XKTB* dataset. This observation prompts additional robustness checks by introducing sample size restrictions on origin prefectures. Specifically, in Figure 1, we progressively tighten the minimum threshold for the total number of out-migrants from origin prefectures, ranging from more than 1 to 30. As the restrictions become stricter, the point estimates of *xiaqu* boundary effects remain stable and statistically significant, albeit with increasing standard errors due to the reduced sample size. In contrast, the point estimates of province boundary effects show a downward trend, becoming statistically insignificant when the threshold exceeds 15 out-migrants. This pattern further supports the observation that the "promoting effects" are largely attributable to prefecture pairs with sparse migration flows.

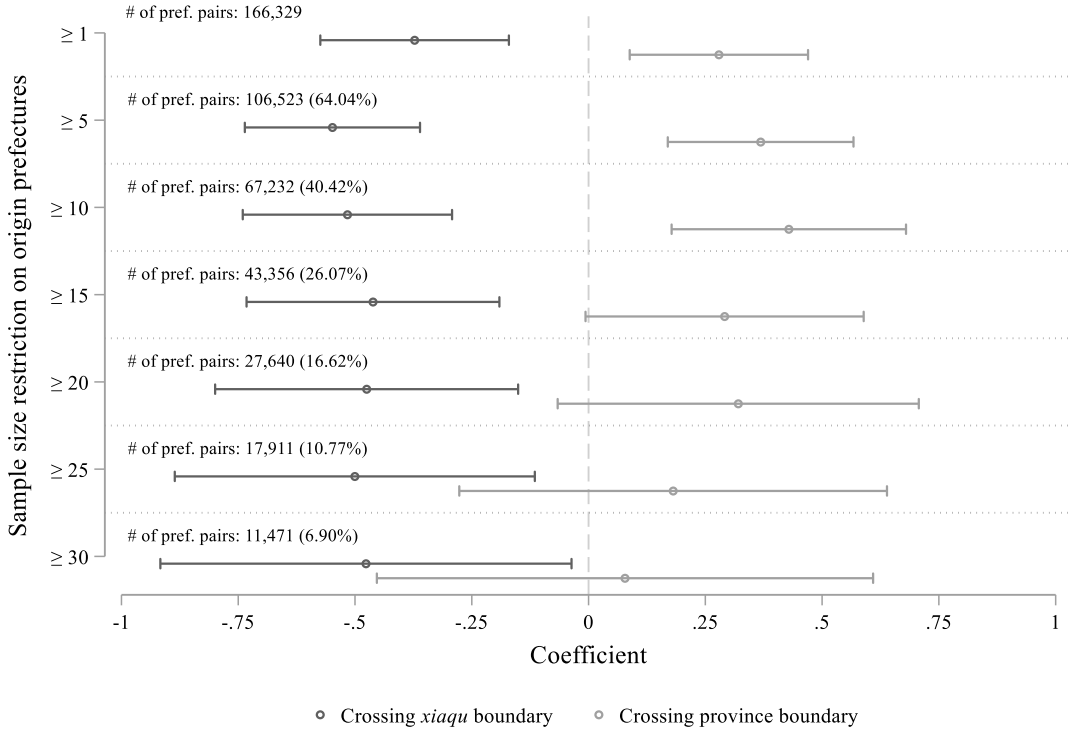

Figure 1 Estimates using sub-samples restricting the number of out-migrants in *XKTB* of origin prefectures
Notes: This figure presents estimates derived from sub-samples, which require origin prefectures to have more than a specified number of out-migrants, as indicated on the y-axis. The estimation specification aligns with that in Column (5) of Table 2. The coefficients for the two variables—indicating whether a migration crossed a *xiaqu* boundary or a provincial boundary—are displayed. Point estimates are represented by dots, while 90% confidence intervals are depicted as lines, calculated using robust standard errors two-way clustering at the origin and destination prefectures. Additionally, the number of effective prefecture pairs included in each estimation is reported, with the fraction relative to the baseline estimation (166,329) provided in parentheses.

Across our baseline estimates and robustness checks, the evidence consistently shows that *xiaqu* boundaries exert a negative effect on migration, suggesting that these higher-level administrative boundaries significantly deter labor mobility. In contrast, the observed "promoting effects" of province boundaries appear to be sensitive to specific outlier pairs. Since we cannot conclude that the patterns observed in these

pairs fully represent randomness, we acknowledge the possibility of mechanisms encouraging migration across province boundaries within the *xiaqu*. However, we caution against over-interpreting these results. A more conservative and robust interpretation is that province boundaries do not deter migration. Given the limitations of our current data, we refrain from making a definitive conclusion between these two perspectives.

Regardless of these interpretations, our findings provide a snapshot of China's vast economy segmented into multiple labor markets by *xiaqu* boundaries. Within these segments, however, administrative barriers to labor mobility are nearly eliminated, facilitating the development of integrated regional labor markets.

Estimation with Province-pair Data and Heterogeneity

This subsection estimates *xiaqu* boundary effects using province pairs rather than prefecture pairs, driven by two key empirical considerations. First, as mentioned earlier, only about 2% of prefecture pairs show positive recorded migrant flows, which may introduce bias. By switching to province pairs, we utilize a data structure where over 40% of pairs show positive flows, providing additional support for our conclusions. More importantly, this approach allows for a more detailed investigation, including estimating heterogeneous boundary effects for each destination *xiaqu* and examining the varying impacts on labor with different skill and entrepreneurship levels. These analyses offer deeper insights into the formation and further influences of *xiaqu* boundary effects.

We construct a province-pair panel dataset using both local residents and migrants from the *XKTB*, encompassing 23 local pairs, 26 inter-province, intra-*xiaqu* pairs, and 480 inter-*xiaqu* pairs. Using the PPML estimator, we estimate the unconditional version of Equation (12), allowing for heterogeneous boundary effects across destination *xiaqu*:

$$m_{od} = \exp[\gamma + \alpha_o + \beta_d + \sigma \ln Distance_{od} + \pi(1 - Culture_{od}) + \sum_x b^x \times \mathbb{I}\{Xiaqu_o \neq Xiaqu_d\} \times \eta_d^x] + \varepsilon_{od}, \quad (13)$$

where η_d^x represents a set of destination *xiaqu* fixed effects. Accordingly, the coefficients $\{b^x\}_x$ capture the average boundary effects of migrating to *xiaqu* x from another *xiaqu*. It is important to note that the term representing province boundary effects is excluded from this estimation equation, as the province-pair dataset is not well-suited for estimating these effects. Specifically, identifying province boundary effects relies on comparisons between fewer than 50 local pairs and inter-province, intra-*xiaqu* pairs, which we find to be highly sensitive to the specification of migration distance.[14] Instead, we treat all intra-*xiaqu* pairs as unaffected by boundary effects and use their comparison with inter-*xiaqu* pairs to identify the *xiaqu* boundary effects.

As we include local residents and local pairs in our estimation to ensure sufficient variation, the estimated *xiaqu* boundary effects also capture part of the disutility associated with leaving one's hometown, which explains the observed larger point estimates. However, since the primary focus of this subsection is on comparing heterogeneous effects, this issue is not expected to have a substantial impact on our analysis, with tentatively assuming that the "leaving-home" effects are similar.[15]

[14] Since individuals remaining within their local province include some inter-prefecture migrants, the average cost due to physical distance is not precisely zero. Referring the approach in Barjamovic et al. (2019), who normalize internal distance to 30 kilometers, we manually assign migration distances for local pairs with values ranging from one kilometer to 30 kilometers, as shown in Appendix Table F.9. While the estimates for *xiaqu* boundary effects remain stable, those for province boundary effects fluctuate significantly. In the main text, we normalize internal distance to one kilometer to ensure that it equals zero after taking the logarithm.

[15] The estimation results of Equation (12) and the aforementioned robustness checks, based on the province-pair dataset, are presented in Appendix Tables F.10 and F.11. Throughout these estimations, consistent patterns are observed.

Figure 2 illustrates the heterogeneous boundary effects across different *xiaqu*. The previous findings are further reinforced by a clear and consistent pattern: *xiaqu* boundaries impose deterrent effects, although the intensity of these effects varies. Specifically, the Shandong *xiaqu* exhibits the most pronounced boundary effects, while the Liangguang *xiaqu* experiences much weaker deterrence.[16]

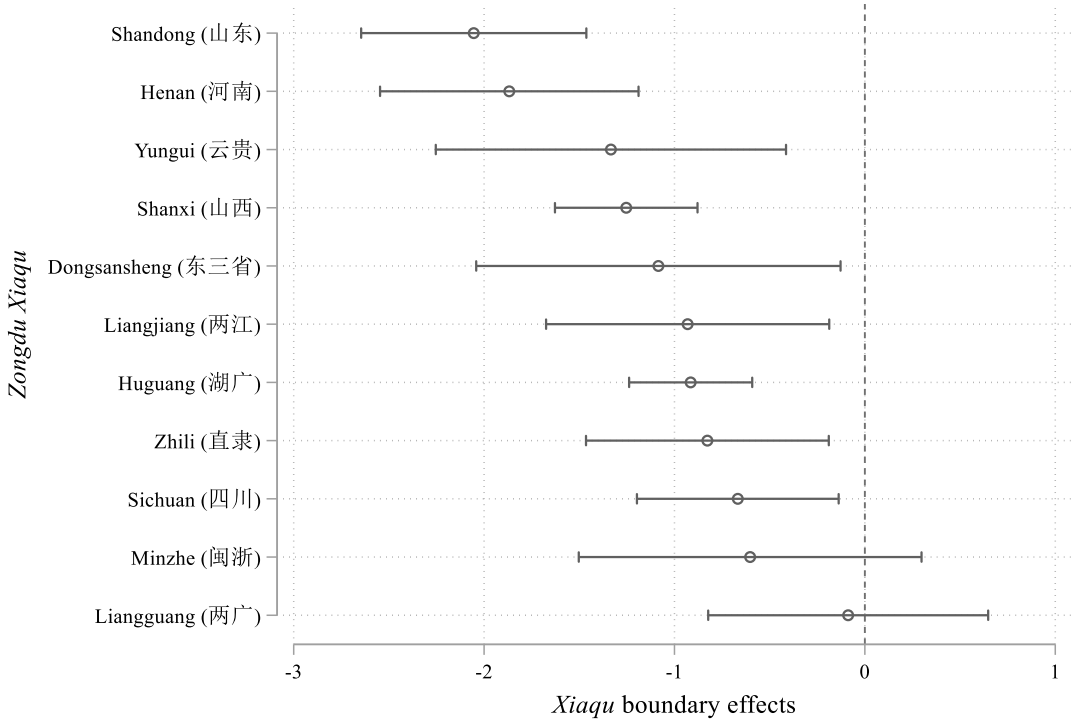

Figure 2 Heterogeneous *xiaqu* boundary effects
Notes: This figure reports the estimated boundary effects for different destination *xiaqu*, derived from estimating Equation (13). The Neimenggu (内蒙古) *xiaqu* is not included in this figure, as there were too few recorded in-migrants to provide accurate estimates. The Shangan (陕甘) *xiaqu* is also excluded, as we observe distinctly large and dubious boundary effects for its third period (see Appendix Table F.13), likely due to data limitations, which result in outlier estimates when combining the three periods. Point estimates are represented by dots, while 90% confidence intervals are depicted as lines, calculated using robust standard errors two-way clustering at the origin and destination provinces.

In the Appendix, we present correlations between the boundary effects and various *xiaqu* characteristics to explore potential insights. Figure F.3 illustrates a negative correlation between *xiaqu* boundary effects and population density, with Shandong *xiaqu* distinctly exhibiting the highest population density. Echoing the narrative of "mainstream migration," which highlights labor outflows to alleviate population pressure on resource exploitation (Ho 1959; Entenmann 1980), the significant population pressure in Shandong *xiaqu* may have contributed to the significant administrative restrictions on migrant inflows. Figure F.4 illustrates a positive correlation between *xiaqu* boundary effects and the degree of trade openness, measured by the proportion of *Haiguan Shangbu* prefectures within each *xiaqu*.[17] This observation aligns with the expectation that trade openness may promote greater openness to migration, potentially offering a partial explanation for the smaller boundary effects observed in the Minzhe *xiaqu* and Liangguang *xiaqu*.

We are also concerned with the heterogeneous restrictions that *xiaqu* boundaries impose on different labor

[16] The table version of these estimates is reported in Appendix Table F.12. Additionally, we present the boundary effects for each *xiaqu* by period in Appendix Figure F.2 and Appendix Table F.13.
[17] *Haiguan Shangbu* (treaty port) refers to a port city that was opened to foreign trade under pressure from Western powers through unequal treaties during the mid-to-late-19th century, where foreign trade drives wider regional market connections.

groups. The first perspective is the occupation structure. In Figure F.5, we observe that *xiaqu* with a larger non-primary sector, particularly a larger tertiary sector, exhibit more significant administrative restrictions on im-migration. This finding contrasts with insights from studies on national grain price correlations, which suggest that the grain retail industry, or even the tertiary sector more broadly, is integrated to a considerable degree in the eighteenth century (Chuan and Kraus 1975; Shiue and Keller 2007; Gu and Kung 2021). Our labor market-based findings reveal significant administrative restrictions in the tertiary sector, suggesting that analyses based on grain prices may offer only a partial perspective. Moreover, given the consensus that the non-agricultural sector is a key driver of economic growth (Herrendorf, Rogerson, and Valentinyi 2014), *xiaqu* boundaries may impede the national industrial transition and hinder the development of agglomeration economies, thereby potentially contributing to the Great Divergence.

Following this logic, we formally investigate the varying resistances of administrative boundaries to workers with different levels of human capital. In Table 4, we divide our migrant sample into subsamples based on skill and entrepreneurship levels, constructing separate province-pair datasets for each subsample.[18] We do not account for heterogeneous effects across destinations in this analysis, as the reduced number of migrants per estimate would not provide sufficient information to accurately capture these heterogeneities. As such, the estimated coefficients reflect the average boundary effects experienced by each group.

In Columns (1) and (2), migrants are categorized as skilled or unskilled. We observe that migration decisions for both groups are deterred by *xiaqu* boundaries, but the point estimates suggest that skilled migrants face less resistance. Additionally, we further divide the migrants into employers and employees, and estimate the effects separately in Columns (3) and (4). We find that both employers and employees experience *xiaqu* boundary effects of similar magnitude.

These findings yield three additional implications. First, *xiaqu* boundaries impede knowledge diffusion. Previous research has highlighted how the incompleteness of formal institutions limits knowledge diffusion (de la Croix, Doepke, and Mokyr 2018) and how certain informal institutions in China, such as clans, can help mitigate these limitations and promote diffusion within smaller spatial regions (Gillette 2016). This study emphasizes that formal institutions themselves can also act as barriers to knowledge diffusion. Second, unskilled workers are more susceptible to these administrative constraints, an issue that persists in modern China (Liang, Song, and Timmins 2024). Administrative forces limit the mobility of unskilled labor within smaller regions, preventing the larger economy from fully capitalizing on its vast pool of low-cost labor. Finally, the mobility of capital appears to have been similarly constrained, as seen in the group of employers. Taken together, these constraints likely impeded large-scale agglomeration at the national level during the Qing Dynasty. In the following sections, we provide further evidence to support this argument.

Table 4 Heterogeneous impacts on labor with different skills and levels of entrepreneurship

	Dependent variable: Migrant share among all individuals from the origin			
	By skills		By levels of entrepreneurship	
	Skilled	Unskilled	Employer	Employee
	(1)	(2)	(3)	(4)
Crossing *xiaqu* boundary	-1.309***	-1.760***	-1.430***	-1.451***
	(0.380)	(0.387)	(0.411)	(0.285)
Origin FEs × Period FEs	Yes	Yes	Yes	Yes
Destination FEs × Period FEs	Yes	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes	Yes
# of individuals in *XKTB*	21,912	11,089	27,186	11,573

[18] We categorize individuals based on their recorded occupation information in *XKTB* reports. Generally, skilled workers refer to those with technical expertise who can perform technical tasks independently or based on instructions or blueprints, while unskilled workers are those who work according to set procedures or guidelines. Employers refer to individuals who own or manage a production unit or assist in managing an enterprise, as well as self-employed individuals. Employees are those who work for an employer and earn a salary. For more details about the categorization, see Appendix B.2.

# of province-pair observations	1,452	1,304	1,542	1,409

Notes: All estimations use the PPML estimator and include constant terms that are not listed in the table. Robust standard errors, with two-way clustering at the origin and destination provinces, are shown in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

LIMITED REGIONAL INTEGRATED LABOR MARKET: THE PERSPECTIVE OF REAL INCOME

In the following two sections, we explore the consequences of the administrative boundary effects on the national labor market, focusing on two key dimensions that serve as fundamental components in analyzing market dynamics: the price, represented by labor income, and the quantity, represented by the spatial allocation of labor.

This section examines how administrative boundaries affect inter-regional labor income divergence. Similar to the law of one price, which suggests that arbitrage reduces price gaps for tradable goods across regions (Parsley and Wei 1996), labor mobility drives real income convergence by reallocating labor supply. For instance, in an integrated national labor market, a positive productivity shock that raises wages in one region attracts labor inflows, increasing supply and moderating wage growth, while labor outflows elsewhere push up wages. This mechanism spreads income shocks nationwide, resulting in more aligned income trends across regions. These dynamics align with findings that economies with highly integrated national labor markets, such as the United States and Japan, exhibit lower levels of regional real income divergence (Li and Lu 2021). In contrast, segmented labor markets weaken correlations between regional income changes, potentially obstructing convergence and exacerbating income disparities.

Our question is whether *xiaqu* boundaries reduce the correlation between real incomes of cross-*xiaqu* location pairs. Addressing this question necessitates overcoming a challenging task of identifying appropriate measures of real income during the Qing Dynasty. While existing estimates offer valuable benchmarks, they encounter several limitations in meeting this specific need. First, many studies estimate real wages using price indices derived from grain prices (for example, Allen 2001; Broadberry and Gupta 2006; Liu 2024); however, given the regulation of grain prices by the government, it is uncertain whether grain price fluctuations reliably capture those of other commodities during the Qing Dynasty. Furthermore, these estimates often focus on specific income sources rather than sufficiently accounting for a broader range of potential income sources, which can occasionally lead to contested interpretations (Deng and O'Brien 2016). Lastly, existing studies face difficulties in producing real income estimates with both nationwide coverage and sufficient spatial granularity, constraining the ability to construct an adequate set of location pairs for investigating real income correlations.

We address these challenges by recovering the real incomes of approximately 250 prefectures and 23 provinces for each sample period, using the structural model developed in the previous section. Our model allows us to infer the average real income from migration patterns. In essence, our estimates leverage these migration patterns to capture the relative attractiveness of each destination, thereby enabling the recovery of real income as the key pull factor. As shown earlier, our theoretical model indicates that the destination fixed effects reflect the average real incomes of destination locations; specifically, $\beta_d = \kappa \ln(\bar{I}_d/P_d)$.[19] Since we can only identify $N-1$ destination fixed effects in the estimation, we center them by subtracting their mean

[19] We acknowledge that the interpretation of our estimated destination fixed effects as real income is contingent on certain assumptions within our model. Specifically, we assume that real income is the sole destination-specific factor influencing migrants' decisions. However, as the spatial economics literature suggests, other factors—such as destination amenities—also play a significant role in shaping migration patterns. If these factors were incorporated into our theoretical model, they would be captured within the destination fixed effects as well. Therefore, a more nuanced interpretation of our recovered real income would be as a measure of overall living standards, which includes not only real income but also amenities and other relevant destination-specific factors that influence migrants' choices.

to derive the centered fixed effects $\{\hat{\beta}_d^c\}_d$, which are given by

$$\hat{\beta}_d^c \equiv \hat{\beta}_d - \frac{1}{N}\sum\nolimits_{d'} \hat{\beta}_{d'} \to_p \kappa \ln\left[\frac{\bar{I}_d/P_d}{\prod_{d'}(\bar{I}_{d'}/P_{d'})^{1/N}}\right]. \quad (14)$$

This equation forms the basis for recovering the relative real income of each prefecture or province.

Recovering the relative real income requires addressing two preliminary steps. The first involves estimating the shape parameter κ, a straightforward task as it appears in Equation (14). According to Equation (12), if accurate data of average real incomes were available, we could replace the destination fixed effects with this data and regress the migrant share on income, alongside other regressors, to estimate κ. While this method is common in the literature (Tombe and Zhu 2019; Bryan and Morten 2019), it is not fully applicable in our case, as the income data is itself what we aim to estimate. As a fallback option, we first rely on the average relative wage estimates of eight provinces provided by Liu (2024) to approximate this parameter. Appendix E provides a detailed account of the data and methodology, yielding a range for κ from 1.64 to 3.89 based on the point estimates from different specifications. In the main text, we use the average of these estimates, which is 2.5.

Our estimates of κ are broadly comparable to those in existing studies. Tombe and Zhu (2019) estimate a similar specification for China's internal migration in 2000 and 2005, obtaining a range of 1.19 to 1.61, while also considering a broader potential range from 1 to 3. For the United States, Fajgelbaum et al. (2019) leverage variations in state-level taxes to estimate it, finding a range of 0.75 to 2.25. Studies in urban economics often report higher estimates, such as 6.83 in Ahlfeldt et al. (2015) and 8 in Dingel and Tintelnot (2020), suggesting modestly dispersed idiosyncratic preferences—likely due to their focus on smaller spatial scales. Overall, our estimates align with the broader range suggested by these diverse strands of research.

The second preliminary step is to assess whether our migrant sample has sufficient power to provide reliable estimates for a series of destination fixed effects. A key concern is that, given the migrant sample size is smaller than the number of prefecture pairs, the data may be insufficient to capture the true attractiveness of each prefecture. Utilizing the province-pair dataset mitigates this concern; however, it introduces another issue akin to that of short-panel data: the limited number of observations for each province in the pair data may result in inconsistent fixed-effect estimates.

In Appendix D.3, we conduct Monte Carlo simulations, constructing simulated migrant samples that mirror the size of our observed data. We set predetermined values for the model parameters close to their estimated values and simulate migration decisions based on the data-generating process outlined in our theoretical model. Across 1,000 simulations, the estimated fixed effects from both prefecture-pair and province-pair data show a strong correlation with the predetermined relative real income. The mean fitted R-squared values between the estimated and true relative real incomes across simulations are 0.695 for the prefecture-pair data and 0.767 for the province-pair data, indicating that, with correct model specifications, our dataset generally reflects the true relative real incomes.

According to Equation (14), we can solve for the relative real income of each prefecture/province compared to the national average. The relative real income for each province, by period, is presented in Appendix Table F.14. Figure 3 compares our estimates for the first sample period, the 1760s, with the wages relative to Zhejiang in eight provinces during 1530 to 1640, as estimated by Liu (2024). Despite the century-long gap, we observe a noticeable positive correlation between these estimates.

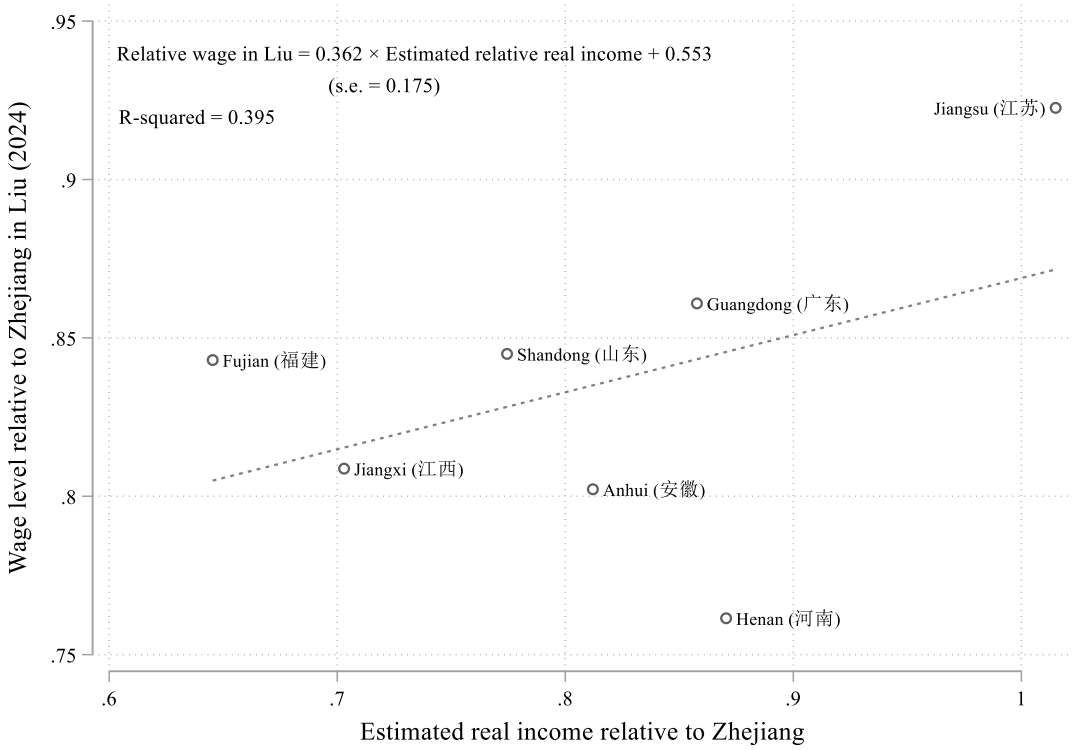

Figure 3 Comparison between our estimated relative real incomes and relative wages in Liu (2024)
Notes: This figure shows the correlation between our estimates of relative real income and the relative wages of eight provinces as estimated by Liu (2024). Our relative real income estimates are derived from Equation (14) using destination fixed effects and a κ value of 2.5. To ensure comparability with Liu's estimates, we normalize each province's relative real income by dividing it by Zhejiang's value, which serves as the reference province with a relative wage of 1. Zhejiang is not shown in the figure, as its value is 1 in both estimations.

Building on the literature that develops price-based approaches to measure market integration (for example, Chen et al. 2024; Fan and Wei 2006; Parsley and Wei 2001), we use the difference in real income changes between two locations to assess their correlation on real income. Specifically, let $\widehat{w}_o$ and $\widehat{w}_d$ represent the real labor income indices for locations o and d, defined as the ratio of current wages to those in the last period, akin to price indices. The logarithmic difference between these indices measures the synchronization of real income changes between the two locations. Based on this, we run the following specification to examine whether prefecture/province pairs across different *xiaqu* exhibit less similarity in their real income change trends:

$$|\ln \widehat{w}_o - \ln \widehat{w}_d| = \alpha + \beta \cdot \mathbb{I}\{Xiaqu_o \neq Xiaqu_d\} + X'_{od}\gamma + \delta_o + \gamma_d + \varepsilon_{od}, \tag{15}$$

where X_{od} controls for the physical distance and cultural differences, δ_o and γ_d represent region fixed effects, and ε_{od} is the error term. We expect to estimate a positive value for β.

Since our recovered relative real incomes reflect income levels for a given period rather than growth, we first need to construct income indices. To do so, we divide our migrant sample from the 1760s into two five-year intervals. We then construct prefecture/province-pair datasets for each of these five-year periods and recover the relative real incomes for 1761–1765 and 1766–1770. By dividing the relative real incomes for each location across these two periods, we obtain the income indices.[20] We apply the same approach to the migrant samples from the 1820s. However, we refrain from doing this with the 1880s sample, as its size is

[20] To recap, what we have recovered are real incomes relative to the national average. Therefore, the income indices represent the change of real income in each prefecture/province relative to the national change. However, the national term is offset when taking the difference between any two locations, so it does not influence the results.

only about half that of the 1760s or 1820s samples individually, and further dividing it would result in too few migrants to estimate reliably.

Table 5 reports the estimated results. Panel A presents estimates using the province-pair data, where both point estimates are positive, with statistical significance observed during the 1760s. Panel B reports estimates based on the prefecture-pair data, incorporating an additional dummy variable to indicate whether two prefectures are located in different provinces. In this case, we observe significant estimates in the second period, including a negative estimate for provincial boundaries in Column (3), which aligns with our narrative of nationally segmented but regionally integrated labor markets. In summary, using real income data inferred from migration patterns, we find suggestive evidence that *xiaqu* boundaries may weaken the correlation between the real incomes of cross-*xiaqu* location pairs, potentially hindering spatial income convergence at the national level.

Table 5 Administrative boundaries and real income correlation

	Dependent variable: Difference in real income changes			
	1761–1770		1821–1830	
	(1)	(2)	(3)	(4)
Panel A. Province-pair data				
Crossing *xiaqu* boundary	0.104***	0.065***	0.032	0.013
	(0.025)	(0.025)	(0.066)	(0.045)
# of province-pair observations	342	342	342	342
Panel B. Prefecture-pair data				
Crossing *xiaqu* boundary	0.004	-0.016	0.039***	0.018‡
	(0.014)	(0.010)	(0.013)	(0.011)
Crossing province boundary	-0.001	0.020	-0.032*	-0.020
	(0.017)	(0.014)	(0.017)	(0.014)
# of prefecture-pair observations	32,220	32,220	31,862	31,862
Origin fixed effects	No	Yes	No	Yes
Destination fixed effects	No	Yes	No	Yes
Controls	No	Yes	No	Yes

Notes: This table explores the impact of administrative boundaries on the spatial correlation of real incomes. The estimations include constant terms, which are not listed in the table. Control variables include physical distance and clan differences between the origin and destination. Robust standard errors are reported in parentheses. ***p < 0.01; **p < 0.05; *p < 0.1; ‡p < 0.11.

LIMITED REGIONAL INTEGRATED LABOR MARKET: THE PERSPECTIVE OF PREFECTURE SIZE DISTRIBUTION

In this section, we focus on the size of each labor markets, specifically, the spatial allocation of labors. We follow a strand of literature that provides empirical evidence for whether Zipf's law holds (Vries 1984; Wrigley 1990). Zipf's law implies that the population of the first largest city is i times that of the i-th largest city (Zipf 1949). While some studies explain Zipf's law as the steady-state outcome of Gibrat's law of proportionate growth (Gabaix 1999; Córdoba 2008), another well-regarded explanation links it to labor mobility and agglomeration economies (Bowen, Munandar, and Viaene 2006; Behrens, Duranton, and Robert-Nicoud 2014). According to the theoretical model developed by Behrens et al. (2014), in an economy with unrestricted labor mobility, sorting based on individuals' utility maximization will lead to a population distribution of cities that aligns with the one predicted by Zipf's law.

Building on this, we use Zipf's law as a benchmark to examine labor spatial distribution and agglomeration across different administrative hierarchies in the Qing Dynasty, aiming to assess the impact of the multilayered administrative boundaries on the efficiency of national labor allocation. Especially, if the sizes of large cities fall short of those predicted by Zipf's law, it may suggest the presence of barriers that hinder spatial agglomeration. For empirical analysis, we adopt the specification used in Gabaix and Ibragimov (2011)

and Barjamovic et al. (2019) by estimating

$$\ln(Rank_i - 0.5) = \alpha - \beta \cdot \ln Population_i + \varepsilon_i, \tag{16}$$

where $Rank_i$ is the population rank of city i, starting from the largest city, and $Population_i$ denotes the population of the i-th largest city. The parameter β represents the Pareto exponent, which would equal one if Zipf's law holds exactly. However, the Pareto exponent has been found to correlate with several economic, geographical, and political factors (Soo 2005), and is highly sensitive to the selection of cities included in the estimation (Rosen and Resnick 1980). An exponent of exactly one is, in fact, more of a "remarkable historical coincidence" (Eeckhout 2004). Therefore, we focus on the goodness of fit to assess alignment with Zipf's law, rather than on the value of the coefficient (similar practice can be found in Anderson and Ge 2005; Li and Lu 2021).

We use the entire prefectural population for the following analysis. While existing studies have shown that defining cities based on urban function rather than administrative division results in a much better fit (Jiang, Yin, and Liu 2015), it is largely infeasible for us to distinguish urban areas from rural areas for each prefecture in the Qing Dynasty. Reassuringly, our analysis focuses on comparing fits across different administrative levels, and we believe this comparison remains informative as long as the division standard is consistently applied.

Figure 4 shows the fit of 296 prefectures across the country in the 1760s. The figure reveals that the size of large cities in the Qing Dynasty was considerably smaller than the fitted line of Zipf's law, indicating that large cities are still not large enough. Notably, the R-squared value of 0.657 is close to a comparable fit for contemporary China.[21] If Zipf's law is interpreted as indicative of sufficient population mobility, this result suggests that national-level labor mobility in the Qing Dynasty approaches levels seen in a modernized country. However, it is important to note that even in today's China, significant institutional barriers to labor mobility, such as the *hukou* system, persist.[22] As previous studies highlight how labor market segmentation in contemporary China constrains economic growth potential (Bosker et al. 2012; Tombe and Zhu 2019), this comparison suggests that the limited national-level labor mobility in the Qing Dynasty similarly hindered its further development.

[21] See sub-figure (a) in Figure 3 of Li and Lu (2021), where the authors investigate the city sizes of 296 administrative cities, measured by the total registered population in 2016. This approach is comparable to our treatment of the entire prefecture as a city. Their fitted plot shows a similar pattern to ours with an R-squared of 0.695.

[22] Established in the 1950s, the *hukou* system is a household registration system initially designed to regulate rural-to-urban migration, which led to significant regional segmentation (Chan and Zhang 1999). Although *hukou* restrictions were gradually eased from the early 1980s, they were not completely eliminated, and migrants continued to face *hukou*-based discrimination in wages and public services well into the 2010s (Song 2014).

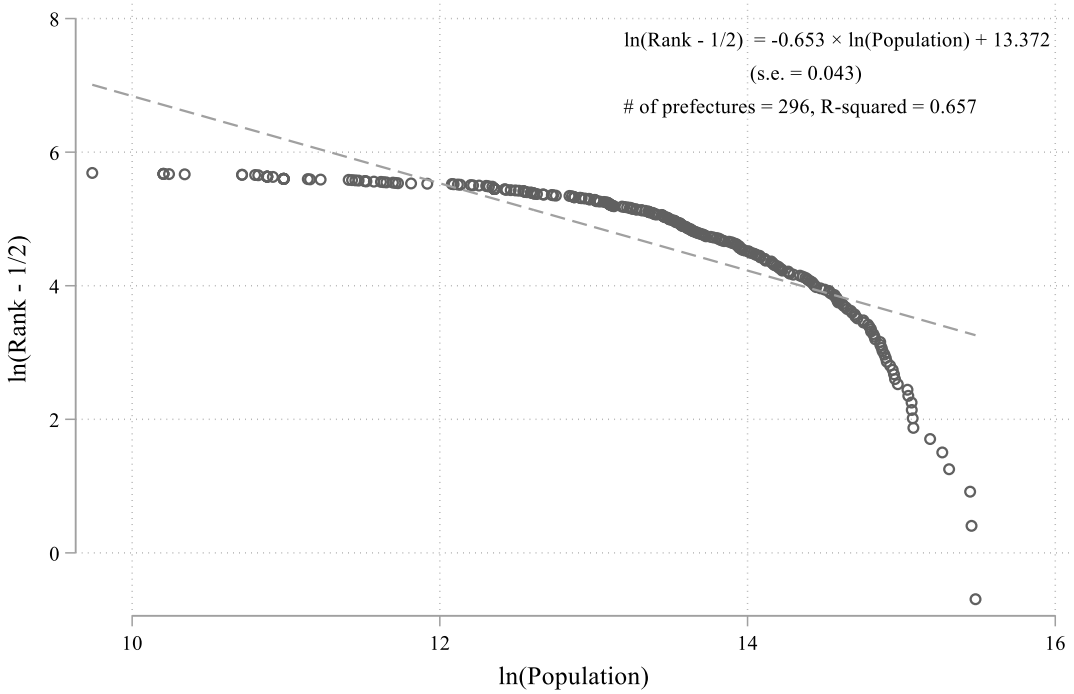

Figure 4 National-level fits of the Zipf's law, 1761–1770
Notes: This figure shows the fits of the Zipf's law at the national level with population data in the first sample period (i.e., 1760s).

Furthermore, we run the fit for prefectures within each *xiaqu* and province, respectively. Table 6 reports the mean R-squared values of fits for seven *xiaqu* containing more than one province and for each individual province. In Panel A, we follow the common approach of ranking and fitting based on population size and observe that the mean R-squared values at both the *xiaqu* and province levels are higher than that at the national level. Notably, we find that the mean goodness of *xiaqu*-level fits is close to that of province-level fits, even with a slightly higher mean observed for *xiaqu*-level fits in the 1760s. This observation suggests that labor mobility within *xiaqu* approaches the levels observed within provinces.

Considering the significant variation in the areas of prefectures, we explore an alternative measure of city size—population density—and recalculate the rank and mean R-squared values, as reported in Panel B of Table 6. In most fits at both the *xiaqu* and provincial levels, we observe a higher goodness-of-fit compared to Panel A. These estimations reveal a larger divergence between the national-level fit and those at sub-national levels. Additionally, *xiaqu*-level fits consistently exhibit better goodness-of-fit than provincial-level fits. These patterns align with and highlight the influence of our earlier findings on multi-layered boundary effects: while soft provincial boundaries impose minimal impediments to migration within a *xiaqu* and even facilitate cross-province migration, *xiaqu* boundaries significantly deter inter-*xiaqu* migration, ultimately constraining the development of mega-cities and national-level agglomeration.

Table 6 The alignment of the Zipf's law at different administrative levels

	Administrative level		
	National level	*Xiaqu* level	Province level
	(1)	(2)	(3)
Panel A. Fit of the number of population and the rank			
Mean of R-squared during 1761 to 1770	0.657	0.732	0.723
Mean of R-squared during 1821 to 1830	0.619	0.722	0.723
Mean of R-squared during 1881 to 1890	0.657	0.693	0.716

Panel B. Fit of population density and the rank			
Mean of R-squared during 1761 to 1770	0.504	0.825	0.740
Mean of R-squared during 1821 to 1830	0.499	0.816	0.747
Mean of R-squared during 1881 to 1890	0.559	0.784	0.715

Notes: This table presents the alignment with Zipf's law at the national, *xiaqu*, and provincial levels, measured by the mean R-squared values obtained from estimating Equation (16) for each location. Panel A fits the logarithm of the rank minus 0.5 against the logarithm of population, while Panel B fits the logarithm of the rank minus 0.5 against the logarithm of population density. For *xiaqu*-level calculations, only *xiaqu* containing more than one province are included. Estimations with fewer than 10 effective data points are excluded from the mean calculation. Detailed results for each estimation are provided in Appendix Tables F.15 to F.17 and are visually presented in Figures F.6 and F.7.

DISCUSSION AND CONCLUSION

Drawing on the rich individual-level migration data in the *Xingke Tiben* (*XKTB*) alongside *Zongdu Xiaqu* and province boundary data in Qing Dynasty, the study employs a structural gravity model and the Poisson Pseudo Maximum Likelihood estimator to evaluate the effects of administrative boundaries on labor migration in late imperial China. *Xiaqu* boundaries are found to significantly deter migration, reducing migrant flows by approximately 31%, while province boundaries exhibit weaker, and occasionally even promoting, effects on inter-provincial migration within the same *xiaqu*. The deterrent effects of *xiaqu* boundaries are shown to vary by region, with stronger impacts in densely populated areas such as Shandong and weaker effects in trade-open regions like Liangguang. These findings illustrate the hierarchical segmentation of labor markets and the significant role administrative divisions played in shaping labor mobility.

Migration data are further used to infer relative real incomes across regions, shedding light on the role of administrative boundaries in limiting income convergence. *Xiaqu* boundaries reduce the correlation of real income changes between regions, indicating that these boundaries inhibit the diffusion of income shocks and perpetuate regional disparities. The analysis demonstrates that income synchronization across cross-*xiaqu* pairs is notably weaker compared to pairs within the same *xiaqu*. These findings highlight the constraining effects of administrative segmentation on national labor market integration, limiting the ability of regional economies to respond collectively to income shifts and economic shocks.

Using Zipf's law to examine the spatial distribution of labor, the study finds that *xiaqu* boundaries disrupt natural patterns of population agglomeration and labor allocation. Regional administrative divisions restrict inter-regional migration toward economic centers, hindering the development of agglomeration economies. These barriers to labor mobility contribute to segmented labor markets, limiting the potential for large-scale economic integration and exacerbating regional imbalances. Such segmentation may have played a critical role in shaping the uneven economic development of late imperial China and offers insights into the historical roots of the Great Divergence.

This study highlights the enduring importance of the state-imposed multilayered administrative boundaries in shaping labor mobility and economic integration. Given China's vast territory, large population, and significant information asymmetries, the central government had to delegate certain powers to local administrations (primarily represented in our study by the *xiaqu* level). This delegation fostered a regional semi-autarkic model characterized by internal integration and external segmentation. For China's long-term economic development, this system was a double-edged sword: while it ensured local social stability and regional market integration, it simultaneously hindered national market integration and the full utilization of labor resources.

This study indicates that the acute principal-agent problem embodied in multilayered administrative boundaries was a primary cause of the lack of national labor market integration in late imperial China. This

market segmentation continues to affect the free movement of labor in present-day China, manifesting in regional policy differences, cultural disparities among other forms of fragmentation. Many countries have similar historical political boundaries that influence contemporary society, making the findings of this study valuable for understanding institutional barriers to market integration on a global scale. In this way, the study represents a unique contribution to global economic history of long-term growth.

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. "The Economics of Density: Evidence From the Berlin Wall." *Econometrica* 83, no. 6 (2015): 2127–2189. https://doi.org/10.3982/ECTA10876

Allen, Robert C. "The Great Divergence in European Wages and Prices from the Middle Ages to the First World War." *Explorations in Economic History* 38, no. 4 (2001): 411–447. https://doi.org/10.1006/exeh.2001.0775

Anagol, Santosh, Fernando V. Ferreira, and Jonah M. Rexer. "Estimating the Economic Value of Zoning Reform." Working Paper Series Working Paper 2021. https://doi.org/10.3386/w29440

Anderson, Gordon, and Ying Ge. "The size distribution of Chinese cities." *Regional Science and Urban Economics* 35, no. 6 (2005): 756–776. https://doi.org/10.1016/j.regsciurbeco.2005.01.003

Bailey, Mark. "Servile Migration and Gender in Late Medieval England: The Evidence of Manorial Court Rolls." *Past & Present* 261, no. 1 (2023): 47–85. https://doi.org/10.1093/pastj/gtac015

Barjamovic, Gojko, Thomas Chaney, Kerem Coşar, and Ali Hortaçsu. "Trade, Merchants, and the Lost Cities of the Bronze Age." *The Quarterly Journal of Economics* 134, no. 3 (2019): 1455–1503. https://doi.org/10.1093/qje/qjz009

Bartz, Kevin, and Nicola Fuchs-Schündeln. "The role of borders, languages, and currencies as obstacles to labor market integration." *European Economic Review* 56, no. 6 (2012): 1148–1163. https://doi.org/10.1016/j.euroecorev.2012.05.008

Bays, Daniel H. "The Nature of Provincial Political Authority in Late Ch'ing Times: Chang Chih-tung in Canton, 1884–1889." *Modern Asian Studies* 4, no. 4 (1970): 325–347. https://doi.org/10.1017/S0026749X00012002

Behrens, Kristian, Gilles Duranton, and Frédéric Robert-Nicoud. "Productive Cities: Sorting, Selection, and Agglomeration." *Journal of Political Economy* 122, no. 3 (2014): 507–553. https://doi.org/10.1086/675534

Bosker, Maarten, Steven Brakman, Harry Garretsen, and Marc Schramm. "Relaxing Hukou: Increased labor mobility and China's economic geography." *Journal of Urban Economics* 72, no. 2 (2012): 252–266. https://doi.org/10.1016/j.jue.2012.06.002

Bowen, Harry P., Haris Munandar, and Jean-Marie Viaene. "Evidence and Implications of Zipf's Law for Integrated Economies." SSRN Scholarly Paper 2006. https://doi.org/10.2139/ssrn.914480

Broadberry, Stephen, Hanhui Guan, and David Daokui Li. "China, Europe, and the Great Divergence: A Study in Historical National Accounting, 980–1850." *The Journal of Economic History* 78, no. 4 (2018): 955–1000. https://doi.org/10.1017/S0022050718000529

Broadberry, Stephen, and Bishnupriya Gupta. "The early modern great divergence: wages, prices and economic development in Europe and Asia, 1500–1800." *The Economic History Review* 59, no. 1 (2006): 2–31. https://doi.org/10.1111/j.1468-0289.2005.00331.x

Brook, Timothy. "The Spatial Structure of Ming Local Administration." *Late Imperial China* 6, no. 1 (1985): 1–55.

Bryan, Gharad, and Melanie Morten. "The Aggregate Productivity Effects of Internal Migration: Evidence from Indonesia." *Journal of Political Economy* 127, no. 5 (2019): 2229–2268. https://doi.org/10.1086/701810

Caliendo, Lorenzo, Fernando Parro, Luca David Opromolla, and Alessandro Sforza. "Goods and Factor Market Integration: A Quantitative Assessment of the EU Enlargement." *Journal of Political Economy* 129, no. 12 (2021): 3491–3545. https://doi.org/10.1086/716560

Campbell, Cameron, and James Lee. "Free and unfree labor in Qing China: Emigration and escape among the bannermen of northeast China, 1789–1909." *The History of the Family* 6, no. 4 (2001): 455–476. https://doi.org/10.1016/S1081-602X(01)00088-4

Cao, Shuji. *Zhongguo Yimin Shi Diliujuan: Qing Shiqi (Migration History of China, Volume 6: Qing Dynasty Republic of China Period)*. Shanghai: Fujian People's Publishing House, 1997.

——. *Zhongguo Renkou Shi: Qing Shiqi (A History of the Chinese Population: The Qing Dynasty)*. Shanghai: Fudan University Press, 2001.

Carpenter, Daniel. "The Evolution of National Bureaucracy in the United States." In *The Executive Branch* 41–71. Oxford University Press , 2005.

Carrieri, Francesca, Vihang Errunza, and Ked Hogan. "Characterizing World Market Integration through Time." *Journal of Financial and Quantitative Analysis* 42, no. 4 (2007): 915–940. https://doi.org/10.1017/S0022109000003446

Casella, Alessandra. "On market integration and the development of institutions: The case of international commercial arbitration." *European Economic Review* The Domain of the State 40, no. 1 (1996): 155–186. https://doi.org/10.1016/0014-2921(95)00044-5

Chan, Kam Wing, and Li Zhang. "The *Hukou* System and Rural-Urban Migration in China: Processes and Changes." *The China Quarterly* 160, (1999): 818–855. https://doi.org/10.1017/S0305741000001351

Chen, Shuo, and James Kai-sing Kung. "Of maize and men: the effect of a New World crop on population and economic growth in China." *Journal of Economic Growth* 21, no. 1 (2016): 71–99. https://doi.org/10.1007/s10887-016-9125-8

Chen, Shuo, Jianan Li, and Qin Yao. "Canal and trade: Transportation infrastructure and market integration in China, 1780–

1911." *Journal of Comparative Economics* (2024). https://doi.org/10.1016/j.jce.2024.08.006
Chilosi, David, Tommy E. Murphy, Roman Studer, and A. Coşkun Tunçer. "Europe's many integrations: Geography and grain markets, 1620–1913." *Explorations in Economic History* 50, no. 1 (2013): 46–68. https://doi.org/10.1016/j.eeh.2012.09.002
Ch'u, T'ung-tsu. *Local Government in China under the Ch'ing*. Cambridge, Mass: Harvard University Asia Center, 1962.
Chuan, Han-sheng, and Richard A. Kraus. *Mid-Ch'ing Rice Markets and Trade: An Essay in Price History*. Harvard University Asia Center, 1975.
Clark, Peter. "Migration in England during the Late Seventeenth and Early Eighteenth Centuries." *Past & Present* 83, no. 1 (1979): 57–90. https://doi.org/10.1093/past/83.1.57
Córdoba, Juan-Carlos. "On the distribution of city sizes." *Journal of Urban Economics* 63, no. 1 (2008): 177–197. https://doi.org/10.1016/j.jue.2007.01.005
Correia, Sergio, Paulo Guimarães, and Tom Zylkin. "Fast Poisson estimation with high-dimensional fixed effects." *The Stata Journal* 20, no. 1 (2020): 95–115. https://doi.org/10.1177/1536867X20909691
Cox, Gary W. "Political Institutions, Economic Liberty, and the Great Divergence." *The Journal of Economic History* 77, no. 3 (2017): 724–755. https://doi.org/10.1017/S0022050717000729
Crespigny, Rafe de. "The Eastern Han." In *Routledge Handbook of Imperial Chinese History* Routledge , 2018.
de la Croix, David, Matthias Doepke, and Joel Mokyr. "Clans, Guilds, and Markets: Apprenticeship Institutions and Growth in the Preindustrial Economy." *The Quarterly Journal of Economics* 133, no. 1 (2018): 1–70. https://doi.org/10.1093/qje/qjx026
Deng, Kent, and Patrick O'Brien. "Establishing statistical foundations of a chronology for the great divergence: a survey and critique of the primary sources for the construction of relative wage levels for Ming–Qing China." *The Economic History Review* 69, no. 4 (2016): 1057–1082. https://doi.org/10.1111/ehr.12281
Dingel, Jonathan I., and Felix Tintelnot. "Spatial Economics for Granular Settings." Working Paper Series Working Paper 2020. https://doi.org/10.3386/w27287
Dobado-González, Rafael, Alfredo García-Hiernaux, and David E. Guerrero. "West versus Far East: early globalization and the great divergence." *Cliometrica* 9, no. 2 (2015): 235–264. https://doi.org/10.1007/s11698-014-0115-9
Dorn, David, and Josef Zweimüller. "Migration and Labor Market Integration in Europe." *Journal of Economic Perspectives* 35, no. 2 (2021): 49–76. https://doi.org/10.1257/jep.35.2.49
Du, Jiaji. "A Study on the Differences of Duty between Zongdu and Xunfu in Qing Dynasty." *Shixue Jikan (Collected Papers of History Studies)* no. 6 (2009): 43–50.
Eeckhout, Jan. "Gibrat's Law for (All) Cities." *American Economic Review* 94, no. 5 (2004): 1429–1451. https://doi.org/10.1257/0002828043052303
Entenmann, Robert. "Sichuan and Qing Migration Policy." *Ch'ing-shih wen-t'i* 4, no. 4 (1980): 35–54.
Evans, Carolyn L. "The Economic Significance of National Border Effects." *American Economic Review* 93, no. 4 (2003): 1291–1312. https://doi.org/10.1257/000282803769206304
Fajgelbaum, Pablo D, Eduardo Morales, Juan Carlos Suárez Serrato, and Owen Zidar. "State Taxes and Spatial Misallocation." *The Review of Economic Studies* (2019). https://doi.org/10.1093/restud/rdy050
Fan, C. Simon, and Xiangdong Wei. "The Law of One Price: Evidence from the Transitional Economy of China." *The Review of Economics and Statistics* 88, no. 4 (2006): 682–697. https://doi.org/10.1162/rest.88.4.682
Findlay, Ronald, and Kevin H. O'Rourke. "Commodity Market Integration, 1500–2000." In *Globalization in Historical Perspective* , edited by Michael D. Bordo, Alan M. Taylor, and Jeffrey G. Williamson, 13–64. University of Chicago Press , 2007. https://doi.org/10.7208/9780226065991-003
Foltz, Jeremy, Yunnan Guo, and Yang Yao. "Lineage networks, urban migration and income inequality: Evidence from rural China." *Journal of Comparative Economics* 48, no. 2 (2020): 465–482. https://doi.org/10.1016/j.jce.2020.03.003
Gabaix, Xavier. "Zipf's Law for Cities: An Explanation." *The Quarterly Journal of Economics* 114, no. 3 (1999): 739–767. https://doi.org/10.1162/003355399556133
Gabaix, Xavier, and Rustam Ibragimov. "Rank − 1 / 2: A Simple Way to Improve the OLS Estimation of Tail Exponents." *Journal of Business & Economic Statistics* 29, no. 1 (2011): 24–39. https://doi.org/10.1198/jbes.2009.06157
Gillette, Maris. *China's Porcelain Capital: The Rise, Fall and Reinvention of Ceramics in Jingdezhen*. Bloomsbury Academic, 2016. https://doi.org/10.5040/9781474259446
Goldstone, Jack A. *Revolution and Rebellion in the Early Modern World: Population Change and State Breakdown in England, France, Turkey, and China,1600-1850; 25th Anniversary Edition*. New York: Routledge, 2016. https://doi.org/10.4324/9781315408620
Gorodnichenko, Yuriy, and Linda L. Tesar. "Border Effect or Country Effect? Seattle May Not Be So Far from Vancouver After All." *American Economic Journal: Macroeconomics* 1, no. 1 (2009): 219–241. https://doi.org/10.1257/mac.1.1.219
Gottschang, Thomas R., and Diana Lary. *Swallows and Settlers: The Great Migration from North China to Manchuria*. University of Michigan Press, 2000. https://doi.org/10.3998/mpub.22808
Graff, David A. "The Reach of the Military: Tang." *Journal of Chinese History* 1, no. 2 (2017): 243–268. https://doi.org/10.1017/jch.2016.35
Granger, C. W. J., and C. M. Elliott. "A Fresh Look at Wheat Prices and Markets in the Eighteenth Century." *The Economic History Review* 20, no. 2 (1967): 257–265. https://doi.org/10.2307/2592156
Greif, Avner, and Guido Tabellini. "The clan and the corporation: Sustaining cooperation in China and Europe." *Journal of Comparative Economics* Institutions and Economic Change 45, no. 1 (2017): 1–35. https://doi.org/10.1016/j.jce.2016.12.003
Gu, Yanfeng, and James Kai-sing Kung. "Malthus Goes to China: The Effect of 'Positive Checks' on Grain Market Development, 1736–1910." *The Journal of Economic History* 81, no. 4 (2021): 1137–1172. https://doi.org/10.1017/S0022050721000437

Guo, Rufei, Junsen Zhang, and Minghai Zhou. “The demography of the great migration in China.” *Journal of Development Economics* 167, (2024): 103235. https://doi.org/10.1016/j.jdeveco.2023.103235
Guy, R. Kent. *Qing Governors and Their Provinces: The Evolution of Territorial Administration in China, 1644-1796*. University of Washington Press, 2013.
Helliwell, John F. “National Borders, Trade and Migration.” *Pacific Economic Review* 2, no. 3 (1997): 165–185. https://doi.org/10.1111/1468-0106.00032
Herrendorf, Berthold, Richard Rogerson, and Ákos Valentinyi. “Chapter 6 - Growth and Structural Transformation.” In *Handbook of Economic Growth* , edited by Philippe Aghion and Steven N. Durlauf, 855–941. Elsevier , 2014. https://doi.org/10.1016/B978-0-444-53540-5.00006-9
Ho, Ping-ti. *Studies on the Population of China, 1368–1953*. Harvard University Press, 1959. https://doi.org/10.4159/harvard.9780674184510
Jacks, D. S. “Market Integration in the North and Baltic Seas, 1500–1800.” *Journal of European Economic History* 33, no. 2 (2004): 285–329.
Jiang, Bin, Junjun Yin, and Qingling Liu. “Zipf’s law for all the natural cities around the world.” *International Journal of Geographical Information Science* 29, no. 3 (2015): 498–522. https://doi.org/10.1080/13658816.2014.988715
Kone, Zovanga L, Maggie Y Liu, Aaditya Mattoo, Caglar Ozden, and Siddharth Sharma. “Internal borders and migration in India.” *Journal of Economic Geography* 18, no. 4 (2018): 729–759. https://doi.org/10.1093/jeg/lbx045
Koss, Daniel. “Political Geography of Empire: Chinese Varieties of Local Government.” *Journal of Asian Studies* 76, no. 1 (2017): 159–184. https://doi.org/10.1017/S0021911816001200
Kuhn, Philip A. *Rebellion and its enemies in late imperial China: militarization and social structure, 1796-1864*. Harvard University Press, 1980.
Kumar, Krishna B., and John G. Matsusaka. “From families to formal contracts: An approach to development.” *Journal of Development Economics* 90, no. 1 (2009): 106–119. https://doi.org/10.1016/j.jdeveco.2008.10.001
Kunt, İ Metin. “Devolution from the Centre to the Periphery: An Overview of Ottoman Provincial Administration.” In *The Dynastic Centre and the Provinces* 30–48. Brill , 2014. https://doi.org/10.1163/9789004272095_004
Leong, Sow-Theng, and G. William Skinner. *Migration and ethnicity in Chinese history: Hakkas, Pengmin, and their neighbors*. Stanford University Press, 1997.
Li, Bozhong, and Jan Luiten van Zanden. “Before the Great Divergence? Comparing the Yangzi Delta and the Netherlands at the Beginning of the Nineteenth Century.” *The Journal of Economic History* 72, no. 4 (2012): 956–989. https://doi.org/10.1017/S0022050712000654
Li, Lillian M. “Integration and Disintegration in North China’s Grain Markets, 1738–1911.” *The Journal of Economic History* 60, no. 3 (2000): 665–699. https://doi.org/10.1017/S0022050700025729
Li, Pengfei, and Ming Lu. “Urban Systems: Understanding and Predicting the Spatial Distribution of China’s Population.” *China & World Economy* 29, no. 4 (2021): 35–62. https://doi.org/10.1111/cwe.12380
Liang, Wenquan, Ran Song, and Christopher Timmins. “Frictional Sorting: The Impacts of Dual Constraints on Mobility and Housing Supply in China.” *International Economic Review* 65, no. 4 (2024): 1747–1776. https://doi.org/10.1111/iere.12724
Liu, Ziang. “Wages, labour markets, and living standards in China, 1530–1840.” *Explorations in Economic History* 92, (2024): 101569. https://doi.org/10.1016/j.eeh.2023.101569
Lorge, Peter. “Military Institutions as a Defining Feature of the Song Dynasty.” *Journal of Chinese History* 1, no. 2 (2017): 269–295. https://doi.org/10.1017/jch.2017.2
Ma, Debin, and Jared Rubin. “The Paradox of Power: Principal-agent problems and administrative capacity in Imperial China (and other absolutist regimes).” *Journal of Comparative Economics* 47, no. 2 (2019): 277–294. https://doi.org/10.1016/j.jce.2019.03.002
McKeown, Adam. “Different transitions: comparing China and Europe, 1600–1900.” *Journal of Global History* 6, no. 2 (2011): 309–319. https://doi.org/10.1017/S1740022811000283
Miller, Gary J. “The Political Evolution Of Principal-Agent Models.” *Annual Review of Political Science* 8, no. Volume 8, 2005 (2005): 203–225. https://doi.org/10.1146/annurev.polisci.8.082103.104840
Mitchener, Kris James, and Mari Ohnuki. “Institutions, Competition, and Capital Market Integration in Japan.” *The Journal of Economic History* 69, no. 1 (2009): 138–171. https://doi.org/10.1017/S0022050709000369
Mostern, Ruth. “Following the Tracks of Yu: Depictions of Imperial Territory.” In *Dividing the Realm in Order to Govern* 57–99. Harvard University Asia Center , 2011. https://doi.org/10.1163/9781684170579_005
North, Douglass C., and Robert Paul Thomas. *The Rise of the Western World: A New Economic History*. Cambridge: Cambridge University Press, 1973. https://doi.org/10.1017/CBO9780511819438
Paker, Meredith M., Judy Z. Stephenson, and Patrick Wallis. “Nominal wage patterns, monopsony, and labour market power in early modern England.” *The Economic History Review* n/a, no. n/a (2024). https://doi.org/10.1111/ehr.13346
Parsley, David C., and Shang-Jin Wei. “Convergence to the Law of One Price Without Trade Barriers or Currency Fluctuations.” *The Quarterly Journal of Economics* 111, no. 4 (1996): 1211–1236. https://doi.org/10.2307/2946713
——. “Explaining the border effect: the role of exchange rate variability, shipping costs, and geography.” *Journal of International Economics* Intranational & International Economics 55, no. 1 (2001): 87–105. https://doi.org/10.1016/S0022-1996(01)00096-4
Peeters, Ludo. “Gravity and Spatial Structure: The Case of Interstate Migration in Mexico.” *Journal of Regional Science* 52, no. 5 (2012): 819–856. https://doi.org/10.1111/j.1467-9787.2012.00770.x
Pomeranz, Kenneth. *The Great Divergence: China, Europe, and the Making of the Modern World Economy*. Princeton, N.J.: Princeton University Press, 2000. https://doi.org/10.2307/j.ctt7sv80
Redding, Stephen J., and Esteban Rossi-Hansberg. “Quantitative Spatial Economics.” *Annual Review of Economics* 9, no. 1 (2017): 21–58. https://doi.org/10.1146/annurev-economics-063016-103713

Richardson, John. “The Administration of the Empire.” In *The Cambridge Ancient History: Volume 9: The Last Age of the Roman Republic, 146–43 BC* , edited by Andrew Lintott, Elizabeth Rawson, and J. A. Crook, 564–598. Cambridge: Cambridge University Press , 1994. https://doi.org/10.1017/CHOL9781139054379.018

Robertson, Raymond. “Wage Shocks and North American Labor-Market Integration.” *American Economic Review* 90, no. 4 (2000): 742–764. https://doi.org/10.1257/aer.90.4.742

Rosen, Kenneth T., and Mitchel Resnick. “The size distribution of cities: An examination of the Pareto law and primacy.” *Journal of Urban Economics* 8, no. 2 (1980): 165–186. https://doi.org/10.1016/0094-1190(80)90043-1

Rowe, William T. *Cities of Jiangnan in Late Imperial China*. New York: State University of New York Press, 1993.

——. “Social Stability and Social Change.” In *The Cambridge History of China* , edited by Willard J. Peterson, 473–562. Cambridge: Cambridge University Press , 2002. https://doi.org/10.1017/CHOL9780521243346.011

Saywell, William, and Raymond Chu. *Career Patterns in the Ch’ing Dynasty: The Office of the Governor General*. Ann Arbor: University of Michigan Press, 2020.

Schlesinger, Jonathan. “Rethinking Qing Manchuria’s Prohibition Policies.” *Journal of Chinese History* 5, no. 2 (2021): 245–262. https://doi.org/10.1017/jch.2020.52

Severen, Christopher. “Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification.” *The Review of Economics and Statistics* 105, no. 5 (2023): 1073–1091. https://doi.org/10.1162/rest_a_01100

Shepherd, John Robert. *Statecraft and Political Economy on the Taiwan Frontier, 1600-1800*. Stanford University Press, 1993.

Shiue, Carol H., and Wolfgang Keller. “Markets in China and Europe on the Eve of the Industrial Revolution.” *American Economic Review* 97, no. 4 (2007): 1189–1216. https://doi.org/10.1257/aer.97.4.1189

Silva, J. M. C. Santos, and Silvana Tenreyro. “The Log of Gravity.” *The Review of Economics and Statistics* 88, no. 4 (2006): 641–658. https://doi.org/10.1162/rest.88.4.641

Skinner, G. William. “Regional urbanization in nineteenth-century China.” In *The city in late imperial China* Stanford University Press , 1977.

Sng, Tuan-Hwee. “Size and dynastic decline: The principal-agent problem in late imperial China, 1700–1850.” *Explorations in Economic History* 54, (2014): 107–127. https://doi.org/10.1016/j.eeh.2014.05.002

Song, Yang. “What should economists know about the current Chinese hukou system?” *China Economic Review* 29, (2014): 200–212. https://doi.org/10.1016/j.chieco.2014.04.012

Soo, Kwok Tong. “Zipf’s Law for cities: a cross-country investigation.” *Regional Science and Urban Economics* 35, no. 3 (2005): 239–263. https://doi.org/10.1016/j.regsciurbeco.2004.04.004

Sotelo, Sebastian. “Practical Aspects of Implementing the Multinomial PML Estimator.” 2019.

Studer, Roman. *The Great Divergence Reconsidered: Europe, India, and the Rise to Global Economic Power*. Cambridge: Cambridge University Press, 2015. https://doi.org/10.1017/CBO9781139104234

Su, Yaqin, Petros Tesfazion, and Zhong Zhao. “Where are the migrants from? Inter- vs. intra-provincial rural-urban migration in China.” *China Economic Review* 47, (2018): 142–155. https://doi.org/10.1016/j.chieco.2017.09.004

Sugihara, Kaoru. “East Asian Path.” *Economic and Political Weekly* 39, no. 34 (2004): 3855–3858.

Tombe, Trevor, and Xiaodong Zhu. “Trade, Migration, and Productivity: A Quantitative Analysis of China.” *The American Economic Review* 109, no. 5 (2019): 1843–1872. https://doi.org/10.1257/aer.20150811

Tsivanidis, Nick. “Evaluating the Impact of Urban Transit Infrastructure: Evidence from Bogotá’s TransMilenio.” (2023).

Vries, Jan De. *European Urbanization 1500-1800*. Methuen, 1984.

Wang, Chenglong, Jianfa Shen, Ye Liu, and Liyue Lin. “Border effect on migrants’ settlement pattern: Evidence from China.” *Habitat International* 136, (2023): 102813. https://doi.org/10.1016/j.habitatint.2023.102813

Wang, Yeh-chien. “Secular Trends of Rice Prices in the Yangzi Delta, 1638-1935.” In *Chinese History in Economic Perspective* , edited by Thomas G. Rawski and Lillian M. Li, 35–68. University of California Press , 1992. https://doi.org/10.1525/9780520377271-006

Wang, Yi. *Transforming Inner Mongolia: Commerce, Migration, and Colonization on the Qing Frontier*. Rowman & Littlefield, 2021.

Wildasin, David E. “Labor-Market Integration, Investment in Risky Human Capital, and Fiscal Competition.” *American Economic Review* 90, no. 1 (2000): 73–95. https://doi.org/10.1257/aer.90.1.73

Williamson, Jeffrey G. “The Evolution of Global Labor Markets since 1830: Background Evidence and Hypotheses.” *Explorations in Economic History* 32, no. 2 (1995): 141–196. https://doi.org/10.1006/exeh.1995.1006

Wong, R. Bin. *China Transformed: Historical Change and the Limits of European Experience*. Cornell University Press, 1997.

Wrigley, E. Anthony. “Urban Growth and Agricultural Change: England and the Continent in the Early Modern Period.” In *The Eighteenth-Century Town* Routledge , 1990.

Yang, Cheng. “A new estimate of Chinese male occupational structure during 1734–1898 by sector, sub-sector pattern, and region.” *The Economic History Review* 75, no. 4 (2022): 1270–1313. https://doi.org/10.1111/ehr.13157

Zelin, Madeleine. *The Merchants of Zigong: Industrial Entrepreneurship in Early Modern China*. Columbia University Press, 2006.

Zhou, Zhenhe, Linxiang Fu, Juan Lin, Yuxue Ren, and Weidong Wang. *Zhongguo Xingzhengquhua Tongshi: Qingdai Juan (A General History of China’s Administrative Divisions: The Qing Dynasty)*. Shanghai: Fudan University Press, 2013.

Zimmermann, Klaus F. “Labour Mobility and the Integration of European Labour Markets.” In *The Integration of European Labour Markets* Edward Elgar Publishing , 2009.

Zipf, George Kingsley. *Human behavior and the principle of least effort*. Human behavior and the principle of least effort.Oxford, England: Addison-Wesley Press, 1949.

For Online Publication

Appendix to:

"Regional Integration within National Segmentation: Multilayered Boundary Effects of Labor Market in Late Imperial China"

A Additional historical background of *Xingke Tiben*

Xingke Tiben (刑科题本, abbreviated as *XKTB*), literally translated as "the court letters (submitted) by the Ministry of Punishment of the Grand Secretariat," are homicide trial records from the Qing Dynasty. In this period, the Emperor held absolute jurisdiction over all homicide cases, irrespective of when, where, or how the homicide occurred. The surviving judicial records of these individual homicide trials comprise 251,857 *XKTB* reports, covering the period from 1734 to 1898. These records were centrally stored in the *Neige* (内阁, the Grand Secretariat) during the Qing Dynasty and are now preserved at The First Historical Archives of China in Beijing.

Political and social stability was a central concern for the Qing Empire's monarchic regime. Consequently, all offenses against the *Daqinglv* (大清律, the National Great Qing Legal Code) that warranted the death penalty, as well as any criminal cases involving unnatural deaths, were required to be reported and processed through a hierarchical judicial system comprising four administrative levels: the county, prefecture, province, and the *Xingbu* (刑部, Ministry of Punishment). According to the *Daqing Huidian* (大清会典, the Qing Empire's regulatory code for officials), senior officials—including county officers, prefectural officers, provincial governors, and Ministry of Punishment officials—were required to submit an initial brief report immediately after a homicide case was brought to court or discovered by government officials. Once the investigation was complete, including the forensic report and the collection of depositions, a more detailed report with sentencing recommendations had to be submitted. This process was mandated regardless of whether all the criminals had been apprehended or the case fully resolved.

When the second report reached the Ministry of Punishment, all the case files—including the depositions from the accuser, the accused, accomplices, and witnesses; the forensic report; the case summary; and sentence recommendations from legal officials across various administrative levels—were compiled into a single, case-specific report. This comprehensive report was then submitted to *Neige*, which presented it to the Emperor for review and a final decision on the sentence. Only after the Emperor's review and approval could the sentence be executed. The finalized report, known as the *XKTB*, was signed and sealed by the Emperor before being archived in the *Neige* archives. The entire process leading to the final sentence is illustrated in Figure A.1.

XKTB reports are classified into two main categories. The first and largest category includes cases related to the death penalty, and the second category consists of autumn-review cases, which document contentious cases requiring additional reviews before being finalized and filed under the first category. There are also other categories, such as prison reports, which do not correspond to individual cases. Our dataset is derived from the first category. Appendix B.1 provides more detailed information on the contents of an *XKTB* report and the process by which we extract relevant data.

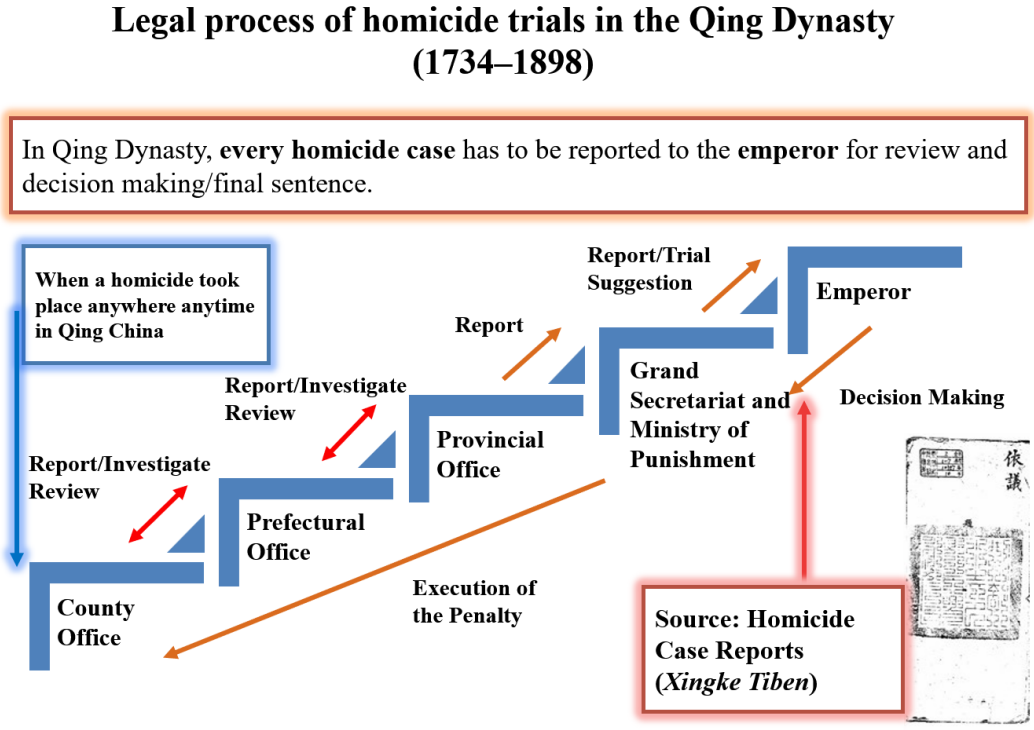

Figure A.1 Legal processing of homicide trials in 1734–1898.

B Data, variable construction, and the discussion of biases

This appendix provides supplementary information regarding our data sources, the methodology used for variable construction, and a discussion of the potential biases of the *XKTB* dataset.

B.1 Data source

As mentioned in the main text, this study relies on three primary data sources: the *XKTB* dataset we constructed, the administrative system of the Qing Dynasty from the Chinese Historical Geographic Information System (CHGIS), and the reconstructed prefectural population data provided by Cao (2001).

XKTB dataset—As introduced in the Appendix A.1, the *XKTB* reports contain information on homicide cases, including extensive demographic and economic details of the accuser, the accused, accomplices, witnesses, and others involved. Our team manually collected approximately 90% of the *XKTB* reports from three time periods—1761–1770, 1821–1830, and 1881–1890—resulting in a total of 52,756 records. We then excluded reports lacking relevant individual information, cases involving government officials (typically embezzlement), and other specific cases, leaving 46,169 records for analysis.

Each *XKTB* report spans approximately 40 pages, featuring six columns with 18 vertically written characters in Classical Chinese. Like most historical texts, the reports lack punctuation between sentences and indentation between paragraphs. Additionally, there are no section titles or consistent markers to separate different parts of the text. Therefore, in contrast to modern census enumerator books, which systematically tabulate data, the structure of the *XKTB* reports is much less uniform. Key information is not consistently located in the same section, requiring the reader to thoroughly examine the entire document to extract relevant data. On average, it takes a trained historian between 45 and 60 minutes to read and extract raw data from a single case.

A typical *XKTB* record normally comprises five parts: the case summary, the forensic report, the depositions, the recommendations for sentences, and the Emperor's decision on the sentence. The summary page typically offers a brief overview of the case, including when, where, and how the homicide occurred. Figure B.1 presents the summary page of one *XKTB* case. In this example, we learn that the incident took place on January 25, 1851 (Lunar Calendar) in Jianchang County (建昌县) in Chengde Prefecture (承德府) of Zhili Province (直隶省) (marked with a green dot on the map we attach in the upper right corner of the figure). The homicide resulted from a dispute over the division of profits from a two-year partnership between two blacksmiths, Liu Yuanxiu (刘沅秀) and Liu Fa (刘发). Liu Yuanxiu, the victim, was a local resident, while Liu Fa, the murderer, was a migrant from Zhangqiu County (章丘县) in Jinan Prefecture (济南府) of Shandong Province (山东省) (marked with a purple dot on the map), situated about 800 kilometers away from Jianchang County, where the crime occurred. Based on this case, we document the details of a local resident from Chengde Prefecture and a migrant from Jinan Prefecture to Chengde Prefecture, including their names, occupations, and other relevant information. The individual-level data, with its fine granularity and extensive geographical and temporal coverage, provides a solid foundation for constructing prefecture- and province-pair datasets, enabling the investigation of the multilayered effects of administrative boundaries.

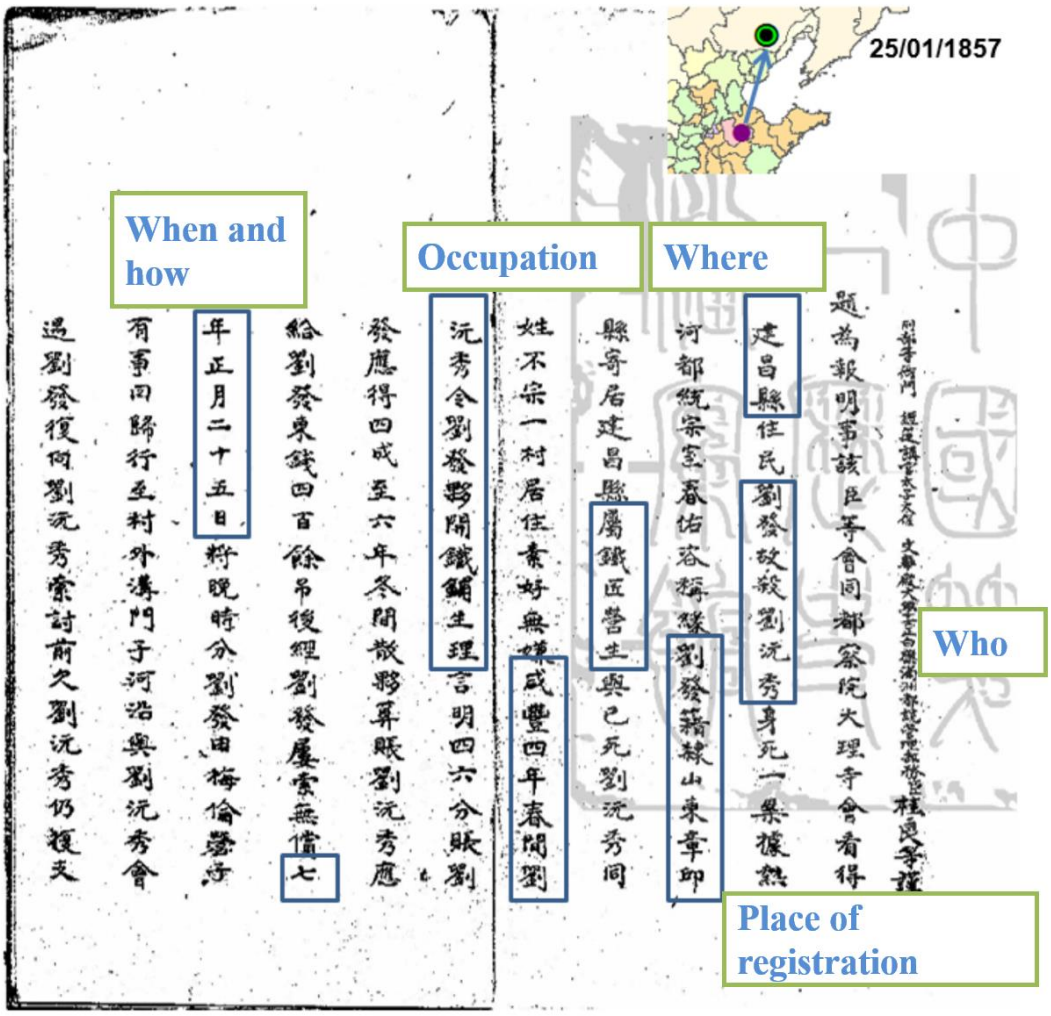

Figure B.1 The summary page of one *XKTB* case.
Notes: This figure is sourced from *XKTB* page dated the 25th of the first Chinese lunar month, 1857. Image courtesy of First Historical Archives of China, FHA 02-01-07-12363-7.

Chinese Historical Geographic Information System (CHGIS)—The CHGIS is a collaborative project between Harvard University and Fudan University, aimed at creating a comprehensive database of populated places and historical administrative units spanning the period of Chinese history from 221 BCE to 1911 CE. During the Qing Dynasty, provinces were the highest administrative units, while prefectures, which grouped several counties together, served as subdivisions within provinces. Starting in 1760, the *Zongdu Xiaqu* became stabilized as quasi-administrative units above the provincial level. By the late eighteenth and nineteenth centuries, both provincial boundaries and *xiaqu* divisions remained largely unchanged, with the exception of the establishment of Taiwan Province from Fujian Province in 1885. Only minor adjustments were made to prefectural boundaries in a few frontier provinces. As such, this administrative structure can be considered stable between 1760 and 1890, the period under examination in our study. Given this historical context, we use the spatial data layers for 1820, provided by the CHGIS Version 6 Dataverse, to identify the administrative boundaries and central points of each prefecture and province.[1] Therefore, Taiwan Province in later periods are treated as part of Fujian Province in our analysis. Additionally, we identify the range of each *xiaqu* based on its encompassing of provinces.

[1] *CHGIS, Version: 6*. (c) Fairbank Center for Chinese Studies of Harvard University and the Center for Historical Geographical Studies at Fudan University, 2016. Available at https://chgis.fas.harvard.edu/data/chgis/v6/ (accessed January 2025).

Prefecture population data—This dataset is derived from Cao's (2001) book *Zhongguo Renkou Shi: Qing Shiqi (A History of the Chinese Population: The Qing Dynasty)*, in which the author estimates the population of each prefecture during the Qing Dynasty by combining direct population data from local sources, such as gazetteers, with population growth rates estimated from anecdotal evidence. We utilize Cao's estimates for three years—1776, 1820, and 1880—and match these with our sample periods: the 1776 estimates correspond to our first sample period (the 1760s), the 1820 estimates to the second sample period (the 1820s), and the 1880 estimates to the third sample period (the 1880s).

The prefecture population data plays two key roles in our study. First, we use it to reweight the count of migrants in *XKTB*, aiming to mitigate the impact of variations in the quantity of homicide trial records across prefectures, which arise from differences in criminal rates, on the accuracy of migrant share calculations. Second, this population data is utilized to examine the spatial distribution of labor and to assess the alignment with Zipf's law across different administrative hierarchies in the main text.

B.2 Variable construction

Migrant share—In the main text, we construct migrant shares as the dependent variable using two methods. The first method relies on the raw counts of migrants and local residents recorded in the *XKTB* dataset. However, the observation rate—measured as the number of individuals recorded in *XKTB* reports per 100,000 people—varies significantly across prefectures, as discussed in detail later. This variation may result in the misinterpretation of locations with lower homicide rates as less attractive for migrants. To address this potential bias, we reweight the counts using the prefectural population estimates provided by Cao (2001).

Taking our first sample period (i.e., 1760s) as the example, we simplify the notation by omitting an additional subscript for the period. Suppose that we observe l_{od}^X migrants from prefecture o to prefecture d and l_{dd}^X local residents in prefecture d recorded in the *XKTB*. The total number of individuals observed in prefecture d in the *XKTB* is then $l_d^X \equiv \sum_{o\in\mathcal{N}} l_{od}^X$. Let l_d^{C} denote Cao's population estimate for prefecture d for this period, then the ratio $r_d \equiv l_d^{\mathrm{C}}/l_d^X$ represents the number of individuals in reality that each recorded individual in the *XKTB* represents.

When constructing the observed migrant share, we use these ratios to reweight the samples. Specifically, the observed share of migrants choosing prefecture d as the destination among all out-migrants originating from prefecture o is calculated by

$$m_{od,-o}^{\mathrm{Data}} = \frac{r_d l_{od}^X}{\sum_{d'\neq o} r_{d'} l_{od'}^X}. \tag{B1}$$

The same method is applied for constructing migrant shares for the second and third periods, after calculating each prefecture's representation ratio using data for the corresponding period. When calculating the province-level migrant share, we still use the prefecture-level representation ratios. For example, the observed share of migrants choosing province j as the destination among both local residents and out-migrants from province i is calculated by

$$m_{ij}^{\mathrm{Data}} = \frac{\sum_{o\in\mathcal{N}_i}\sum_{d\in\mathcal{N}_j} r_d l_{od}^X}{\sum_{o\in\mathcal{N}_i}\sum_{d\in\mathcal{N}} r_d l_{od}^X}, \tag{B2}$$

where $\mathcal{N}_i$ and $\mathcal{N}_j$ denote the sets of prefectures in provinces i and j, respectively.

Migration distance—To measure migration distance, we follow the approach of Barjamovic et al. (2019) in estimating the gravity model, using the linear distance (i.e., the shortest path over the Earth's surface)

between two points. Migration distance is calculated using the Haversine formula, which accounts for the Earth's spherical shape. Let the latitudes of the central locations of the origin and destination in radians be denoted as φ_o and φ_d respectively, and their longitudes in radians as λ_o and λ_d. The formula is

$$Distance_{od} = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\varphi_{od}}{2}\right) + \cos(\varphi_1) \cdot \cos(\varphi_2) \cdot \sin^2\left(\frac{\Delta\lambda_{od}}{2}\right)}\right), \tag{B3}$$

where r is the radius of the Earth (commonly approximated as 6,371 km), and $\Delta\varphi_{od} = \varphi_o - \varphi_d$ and $\Delta\lambda_{od} = \lambda_o - \lambda_d$ are their differences in latitude and longitude, respectively.

Clan difference—This variable is constructed based on the difference in surname composition between any two prefectures/provinces. The procedure is as follows:

1. Exclude all individual samples without name records, leaving 99.67% (125,945/126,366) of our samples across the three periods. Then, extract the surname of each individual by taking the first character of their name, resulting in 1,294 distinct surnames.
2. Specify an order for each surname
$$[Surname_1, Surname_2, \dots, Surname_{1294}].$$
3. Match individual samples to each prefecture/province o if the individual is a local resident of o or migrates from o during the corresponding sample period. These matched samples are used to represent the ex-ante surname compositions. With the matched samples, calculate $N_o^{Surname_i}$, which denotes the number of individuals with $Surname_i$. Accordingly, we construct a surname vector for each prefecture/province:
$$SurnameVector_o \equiv \left[N_o^{Surname_1}, N_o^{Surname_2}, \dots, N_o^{Surname_{1294}}\right].$$
4. Construct the prefecture/province pairs and calculate the cosine similarity between the surname vectors of the two prefectures/provinces in each pair:
$$Culture_{od} = \frac{SurnameVector_o \cdot SurnameVector_d}{\|SurnameVector_o\| \times \|SurnameVector_d\|}. \tag{B4}$$

We illustrate the similarity in clan compositions between each pair of provinces in Figure B.2, where a "warmer" cell indicates a higher degree of similarity in the clan compositions of the two corresponding provinces. The measure of clan difference between each pair of prefectures is calculated using the same method but is not presented here due to the large number of pairs.

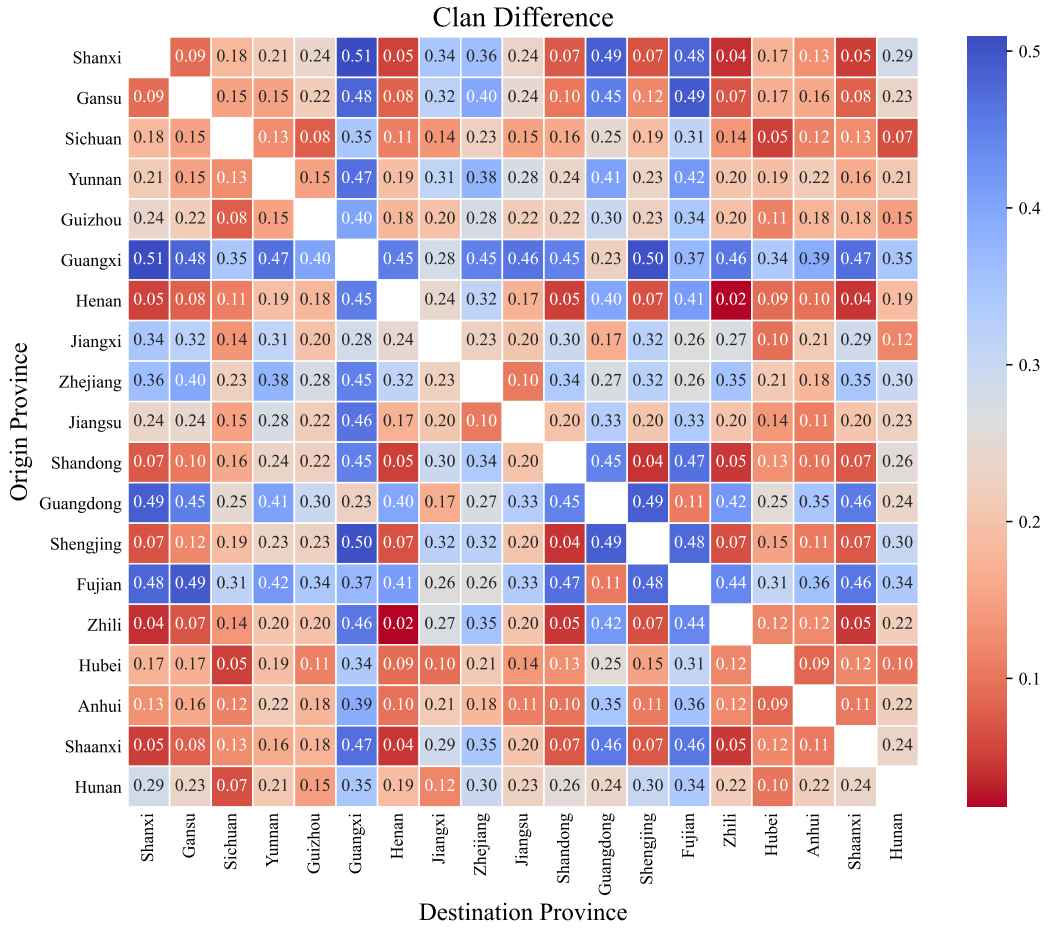

Figure B.2 Surname difference between each two provinces.
Notes: This heatmap illustrates the clan culture differences between each pair of provinces, calculated by $1 - Culture_{od}$. A "warmer" cell indicates greater similarity (i.e., less difference) in the clan compositions between the two corresponding provinces.

Skilled or unskilled worker—In the main text, we categorize migrants based on skills and entrepreneurship to examine heterogeneous *xiaqu* boundary effects on groups with varying levels of human capital. As mentioned earlier, the *XKTB* reports lack uniform coding for occupations, necessitating manual extraction of occupation data and classification for each individual. To avoid creating overly fine distinctions when collecting raw data, we group individuals into three broad categories: skilled workers, semi-skilled workers, and unskilled workers. Generally, skilled workers are those with technical expertise who can perform technical tasks independently; semi-skilled workers have some technical skills but require instructions or blueprints to operate; and unskilled workers have no specialized technical knowledge and work according to set procedures or guidelines. The criteria for categorization are industry-specific, detailed as follows:

- *Primary sector:* Skilled workers primarily include individuals with specialized skills in fields like agricultural production, fisheries, and forestry. This category includes farmers knowledgeable in using agricultural tools and pesticides, as well as ship captains. Semi-skilled workers are production workers who rely on instructions or blueprints to perform tasks, such as assistants on livestock farms or villagers operating equipment. Unskilled workers are typically engaged in purely physical labor, such as manual work on farms or in fishing, without the need for specialized skills or technical knowledge.
- *Secondary sector:* Skilled workers primarily include individuals with specialized technical skills, such

as those in mechatronics, sheet metal work, welding, and riveting. This category includes workers involved in machine processing in the machinery manufacturing industry, welders, and similar roles. Semi-skilled workers are production workers who rely on instructions or blueprints to perform tasks, such as assemblers or operators on assembly lines. Unskilled workers mainly engage in purely physical labor, such as cleaners, decontaminators, and similar roles that do not require specialized technical knowledge.

- *Tertiary sector:* Skilled workers primarily include those with specialized service skills, such as beauticians proficient in various makeup techniques and chefs skilled in diverse cooking methods. Semi-skilled workers are those who follow operational guidelines, such as processes and steps, for tasks like customer service representatives and purchasing assistants. Unskilled workers, on the other hand, are typically those who do not require technical expertise, such as security guards and cleaners.

In Table 4, we combine skilled workers and semi-skilled workers into a single category labeled "skilled," while unskilled workers are grouped under the label "unskilled."

Employer or employee—We also categorize individuals into three levels of entrepreneurship: employers, self-employed individuals, and employees. Employers are those who own or manage a production unit, or who assist in managing the enterprise. Self-employed individuals are those who organize economic activities either as an individual or within a family unit. Employees are those who work for an employer and earn a salary. In Table 4, both employers and self-employed individuals are grouped under the label "employer," while employees are categorized as "employee."

B.3 Potential biases of the XKTB dataset

Although the *XKTB* dataset offers a considerable number of individual-level samples with sufficient spatial granularity and coverage, it may contain inherent biases due to its exclusive focus on homicide cases. In this section, we endeavor to provide a comprehensive discussion of these potential biases and present additional supporting evidence to assess its reliability.

The first potential bias arises from the imbalanced geographic distribution of homicide cases. Although the dataset covers most of China, excluding its frontier regions, the observation rate (measured as the number of individuals recorded in *XKTB* reports per 100,000 people) varies significantly across prefectures (Yang 2022). This rate tends to be negatively correlated with factors such as the degree of imperial control, economic development, and population density. If the migrant share is naively calculated using the raw counts from the *XKTB*, locations with lower homicide rates may be inaccurately perceived as less attractive in the estimation. To address this issue, as discussed in Appendix B.2, we first calculate how many individuals each recorded in the *XKTB* represents for each prefecture, based on population estimates from Cao (2001). We then reweight the migrant and local resident counts by these ratios when calculating migrant shares, which helps mitigate the geographic bias.

The second potential bias concerns the sample selection. Although homicide records from other countries have also been used to investigate migration patterns, particularly in contexts with limited historical data (Clark 1979; Bailey 2023), relying on criminal samples may introduce inherent limitations in representing the broader population. A general justification is that, as mentioned earlier, individuals recorded in the *XKTB* reports include not only criminals but also bystanders, such as witnesses and accusers, who are likely more uniformly distributed across the population. The following discussion further addresses four aspects of these concerns:

Firstly, it is often assumed that poorer and/or lower-class individuals are more likely to be involved in homicides than wealthier individuals, which could introduce bias into migrant rates if migration correlates with wealth or social class. There are two key arguments addressing this concern. First, while individual income and class data are unavailable, Yang (2022) finds that individuals involved in homicides—whether as murderers, victims, or bystanders (such as witnesses or accusers)—show only modest differences in their occupational structures, particularly in the proportion of agricultural versus non-agricultural workers. This suggests that wealth and class bias may be relatively small. Second, even if such bias exists, our analysis, based on prefecture-pair data, excludes local pairs and focuses on comparisons between intra-province migrants, inter-province, intra-*xiaqu* migrants, and inter-*xiaqu* migrants. Therefore, as long as the sample bias is relatively consistent across these groups, our data should still provide reliable estimates.

Secondly, individuals involved in homicide cases tend to be between the ages of 25 and 50, which reduces the representativeness of the sample for children and the elderly. This selection is also shown in Figure B.3, where we find that the mean age (36.7) of migrants recorded in the *XKTB* reports is larger and the distribution is more concentrated, compared to the age population calculated by Banner population in rural Liaoning constructed by Lee and Campbell (2016). Despite the existence of this bias, since this group plays a dominant role in the labor supply, they are more likely to align with the sorting patterns based on real income and migration costs, as described in our theoretical model and the gravity equation. Therefore, from the perspective of estimating boundary effects, this sample selection may have actually filtered in individuals who are more representative of what we aim to estimate.

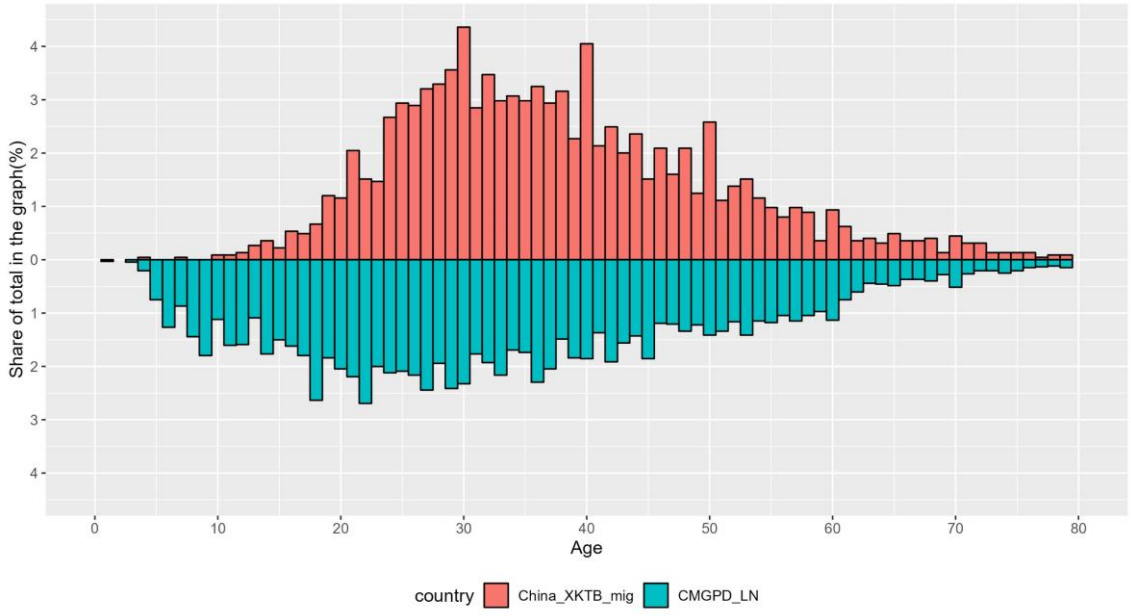

Figure B.3 A comparison of age distribution between migrants recorded in the *XKTB* reports and the Banner population.
Notes: This figure compares the age distribution of migrants recorded in the *XKTB* reports with that of the Banner population in rural Liaoning, as constructed by Lee and Campbell (2016). The red bins represent the age distribution based on the *XKTB* reports, while the blue bins represent the age distribution of the Banner population.

Thirdly, there is a significant gender imbalance in the *XKTB* dataset, with approximately 90% of the recorded individuals being male. This imbalance is common in historical materials. While females played a significant role in society, most historical records from pre-modern China disproportionately document males. To maintain consistency, we use only the male sample when constructing the prefecture-pair and province-pair datasets. However, this observation highlights a potential source for female individual-level data. We are

optimistic that, with further collection of *XKTB* reports, the female sample will eventually reach a sufficient size for future studies focused on women. Regarding our estimation of boundary effects, we conducted a simple comparison of migration distances between males and females. As shown in Figure B.4, we found only a slightly shorter mean migration distance for females, which reassures us that gender differences in migration are likely not significant. Additionally, although males tend to migrate longer distances and cross boundaries more frequently than females, this could lead to an overestimation of cross-boundary migrant shares and potentially an underestimation of the boundary effects. In summary, we believe the gender selection introduces minimal bias or, at most, results in slightly conservative estimates.

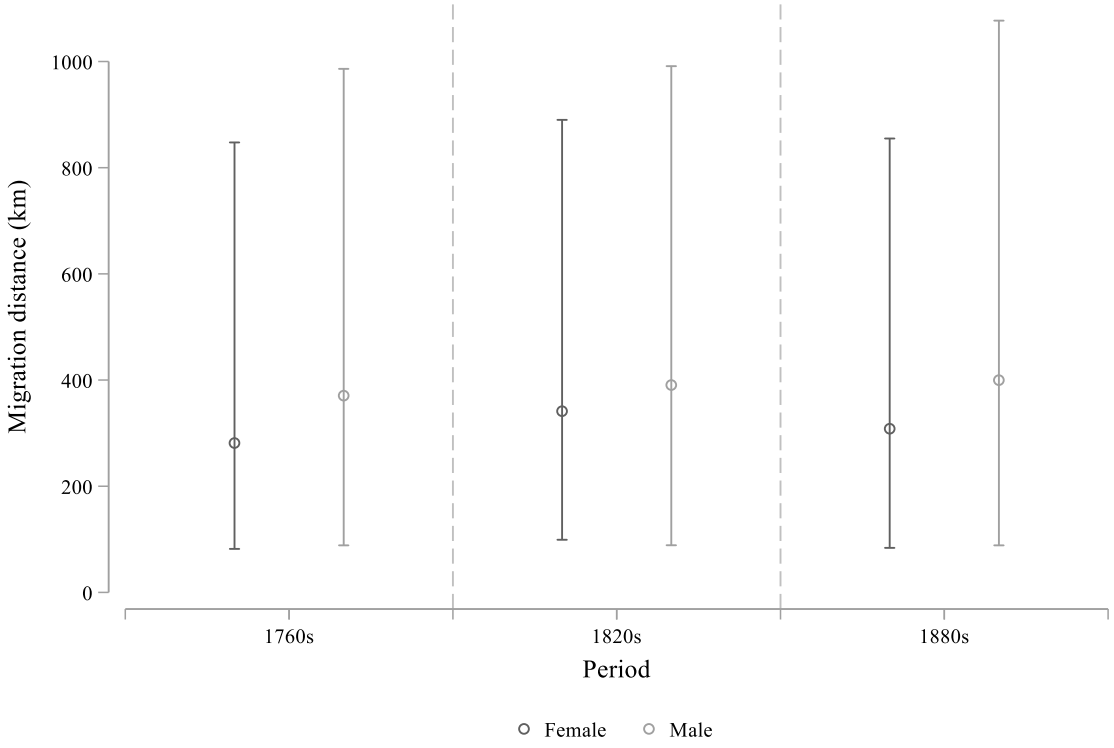

Figure B.4 Migration distance by gender.
Notes: This figure compares migration distances between males and females. The dots represent the mean migration distance, while the intervals indicate the range between the 5th and 95th percentiles.

Fourthly, one might be concerned that the composition of victims, perpetrators, and other individuals involved in homicide cases could vary across regions. If these groups demonstrate systematically different migration preferences, such regional variations in composition could introduce biases when comparing different location pairs to estimate boundary effects. To address this concern, we calculate the proportion of murderers, victims, and others involved in each homicide case and then compute the mean proportion for each prefecture. Table B.1 summarizes these means. The average proportion of perpetrators across prefectures in the 1760s is 0.425, with a relatively small standard deviation of 0.076, indicating that the composition of homicide cases is comparable across prefectures, a pattern also observed for other proportions in each period. Similar to other proportions in each period. Furthermore, Figure B.5 graphically presents the mean proportion of murderers in the 1760s by prefecture, again showing that the mean proportions are relatively consistent across different regions. These results suggest that the impact of this concern on our estimates may not be significant.

Table B.1 The mean proportion of murders, victims, and others involved in a single homicide case across prefectures.

	1761–1770	1821–1830	1881–1890
Murderers	0.425	0.457	0.494
	(0.076)	(0.066)	(0.070)
Victims	0.298	0.333	0.371
	(0.072)	(0.058)	(0.074)
Others involved	0.277	0.210	0.135
	(0.084)	(0.073)	(0.073)

Notes: This table summarizes the proportion of murderers, victims, and other individuals involved in a single homicide case across three sample periods. The means for all prefectures are presented, with standard deviations reported in parentheses.

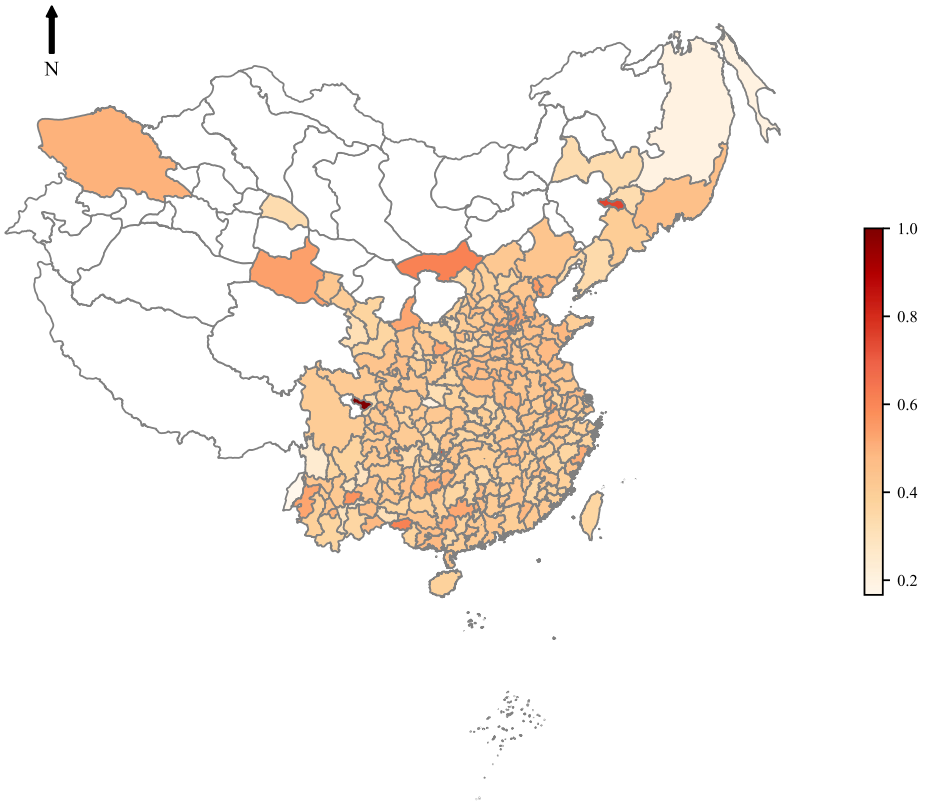

Figure B.5 The mean proportion of murderers in a single case in the 1760s.

The third potential bias concerns the possible deterioration in the quality of homicide records over time. This concern is not unfounded, as we observe a decrease in the text length of the *XKTB* reports during the third sample period (i.e., the 1880s), alongside a declining trend in the mean number of individuals recorded in a single report—dropping from 3 in the 1760s to 2.73 in the 1820s and further to 2.42 in the 1880s. There are at least three possible explanations for this observation. First, national or large-scale regional shocks, such as the First Opium War (1840–1842), the Taiping Rebellion (1851–1864), the course change of the Yellow River in 1855, and the Dingwu Qihuang (the Northern Chinese Famine of 1876–1879), caused significant population loss and weakened the government's administrative power (Cao 2001). Second, the observed reduction in text length might reflect an effort to increase efficiency by producing more concise finalized reports, particularly in cases where the evidence was sufficiently clear.[2] Third, this trend could be attributed to societal advancements and improvements in public security, which likely reduced the overall homicide

[2] This is supported by a comparison between the provincial report of a homicide trial, "FO931 242," archived at the National Archives in London, and the finalized report, "02-01-07-3516-001," stored at the First Historical Archive in China. In the final stage of the legal process, the suspect was confirmed as the murderer, while the other individuals were proven innocent. Consequently, detailed information about the witnesses and the accuser was removed in the finalized *XKTB* report.

rate (Chen, Peng, and Zhu 2017).[3] There are two arguments to address this concern on our estimation. First, regardless of the underlying cause, these shocks were likely national in scope. Since our estimation of boundary effects relies on comparisons among different location pairs, we expect that such national shocks would not significantly bias our estimates, as their influence is unlikely to systematically target specific location pairs. Second, in the main text, we present estimates for each of the three periods individually and observe consistent qualitative results across these periods.

To further assess the reliability of the *XKTB* data, we compare several measures derived from it with existing benchmarks. The first comparison examines the in-migrant shares (the proportion of in-migrants in the total population) of specific regions, as estimated from the *XKTB* dataset, against those estimated by Cao (2001). Figure B.6 presents this comparison, showing that the in-migrant rates estimated from the *XKTB* dataset closely align with Cao's estimates in most regions, with a maximum discrepancy of 25%. For instance, in the case of Ningguo Prefecture, an area heavily affected by the Taiping Rebellion, one possible explanation for this discrepancy is that Cao's estimate of the natural growth rate—based on the 1880–1953 growth rate—might be lower than the actual natural growth rate. This underestimation is plausible, as rising fertility typically follows significant population losses due to war. Cao's methodology relies on the recorded population in 1953 and an assumed natural growth rate of 7% to backcast the population for 1889. By comparing this reconstructed population with the recorded number of local residents, in-migration rates are derived. Therefore, underestimating the natural growth rate would lead to an overestimation of in-migration rates. Overall, our estimates are generally lower than Cao's, making them conservative and potentially serving as lower-bound estimates.

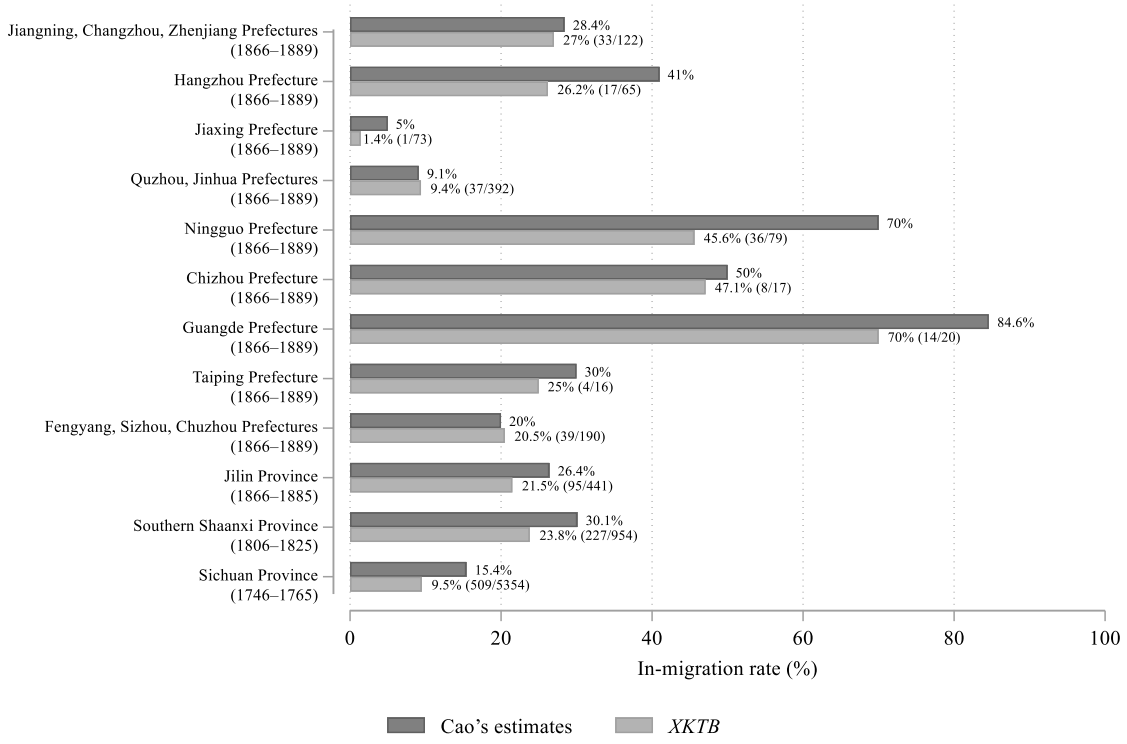

Figure B.6 A comparison of in-migrant shares in specific regions between the *XKTB* dataset and Cao's estimates.
Notes: This figures compared in-migrant shares estimated based on the *XKTB* dataset with estimates in Cao's (2001) book *Zhongguo Renkou Shi: Qing Shiqi (A History of the Chinese Population: The Qing Dynasty)*.

[3] In Chen, Peng, and Zhu (2017), the number of homicide cases is used as a measure for the homicide rate, implicitly interpreting a decrease in the number of homicide cases as a decline in the homicide rate.

Additionally, we compared the distribution of the origin regions of in-migrants in Taiwan, as estimated by *XKTB*, with Cao's estimates. As shown in Figure B.7, we again observed a high degree of alignment. These results further bolster our confidence in the reliability of the data.

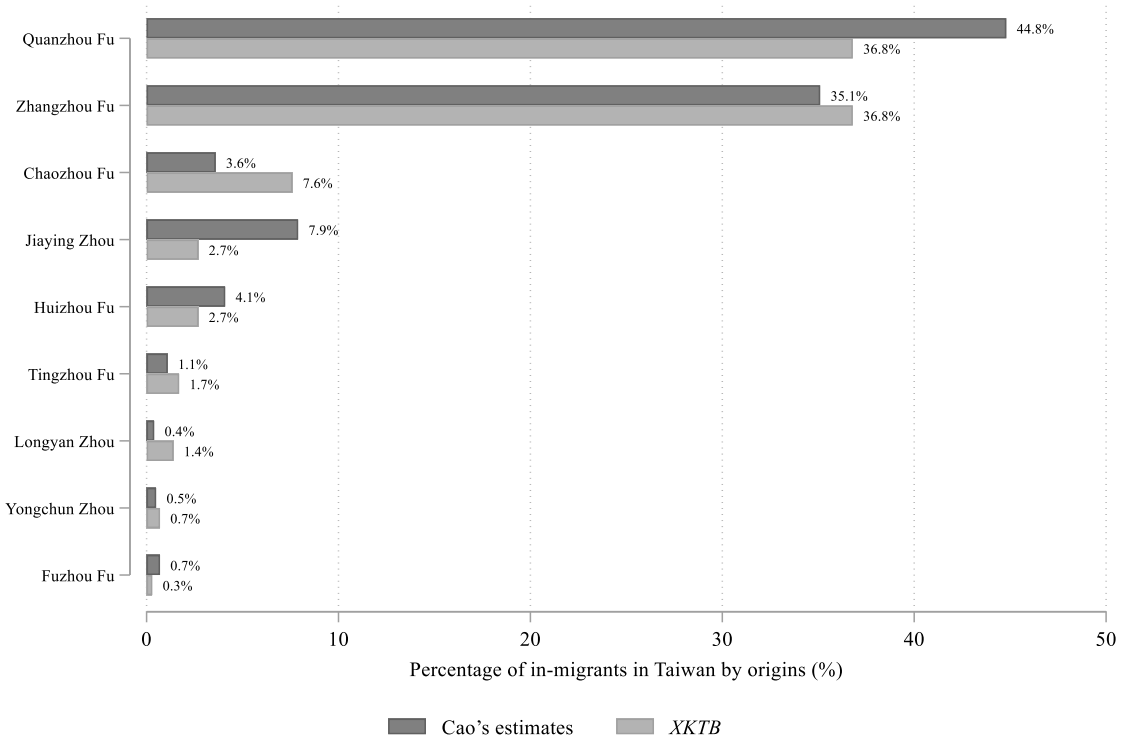

Figure B.7 A comparison of the in-migrant origin prefecture distribution in Taiwan between *XKTB* and Cao's estimates.
Notes: This figure compares the in-migrant origin prefecture distribution in Taiwan estimated from the *XKTB* dataset with the distribution presented in Cao's (2001) book *Zhongguo Renkou Shi: Qing Shiqi (A History of the Chinese Population: The Qing Dynasty)*.

Finally, we compare our estimates with those derived from genealogical records. Liu (1983), in the analysis of the Wei family genealogy in Hengyang, found that the out-migrant share (i.e., the proportion of out-migrants) among individuals in the 19th to 21st generations of the family (birth years between 1704 and 1795) was 10.89%. This share is remarkably close to the out-migrant share calculated from all *XKTB*-recorded individuals in Hengzhou Prefecture during the 1760s, which stands at 10.49%. Additionally, we derived the destination province distribution for out-migrants from Hengzhou Prefecture based on Liu's (1983) findings and similarly constructed the distribution using the *XKTB* reports. As shown in Figure B.8, the two distributions are generally comparable, with the exception of a notable discrepancy observed in Hubei Province.

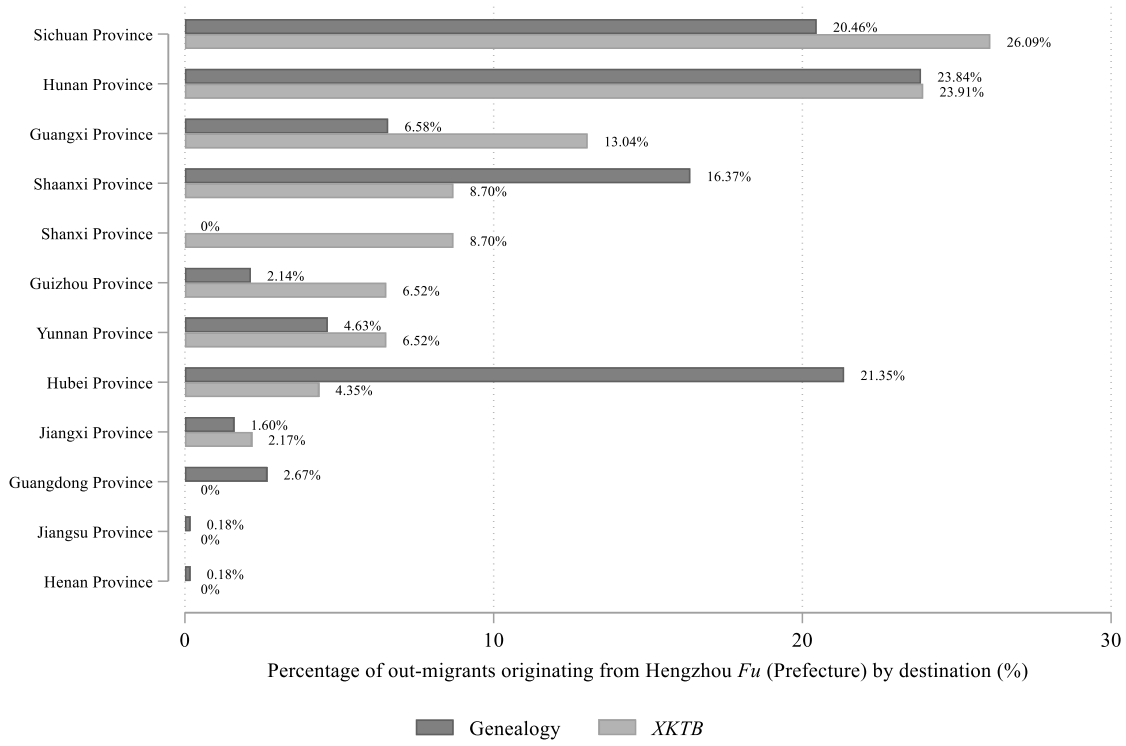

Figure B.8 A comparison of the out-migrant destination distribution originating from Hengyang *Fu* (Prefecture) between *XKTB* and genealogy.
Notes: This figure compares the out-migrant destination distribution originating from Hengyang *Fu* (Prefecture) as derived from the *XKTB* reports and genealogical records. The latter distribution is constructed based on Liu (1983).

C Proofs of the structural model

C.1 Proof of Equation (5)

Before determining the optimal migration decisions of individuals, we must first solve for the optimal utility associated with each destination. Since μ_{od} and ϵ_{od}^{i} are both positive and predetermined for a given destination d, the utility optimization problem reduces to selecting the optimal consumption of each variety of commodities. Formally, the problem can be expressed as:

$$\max_{\{c_{od}^{i}(\omega)\}_{\omega\in\Omega}} u = \left[\sum\nolimits_{\omega\in\Omega} c_{od}^{i}(\omega)^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}},$$

$$\text{s.t.} \sum\nolimits_{\omega\in\Omega} c_{od}^{i}(\omega)p_d(\omega) \le I_d^i \text{ and } c_{od}^{i}(\omega) \ge 0.$$

Since the marginal utility of commodity ω is

$$u_\omega = \left[\sum\nolimits_{\omega\in\Omega} c_{od}^{i}(\omega)^{\frac{\eta-1}{\eta}}\right]^{\frac{1}{\eta-1}} c_{od}^{i}(\omega)^{-\frac{1}{\eta}} > 0, \text{ for } c_{od}^{i}(\omega) > 0 \tag{C1}$$

and

$$\lim_{c_{od}^{i}(\omega)\to 0} u_\omega = +\infty, \tag{C2}$$

utility optimization necessitates that individuals consume all commodities and allocate their entire income to consumption. In other words, we have $\sum_{\omega\in\Omega} c_{od}^{i}(\omega)p_d(\omega) = I_d^i$ and $c_{od}^{i}(\omega) > 0$, $\forall\omega$ (no corner solution).

Accordingly, the Lagrangian is

$$\mathcal{L} \equiv \left[\sum\nolimits_{\omega\in\Omega} c_{od}^{i}(\omega)^{\frac{\eta-1}{\eta}}\right]^{\frac{\eta}{\eta-1}} - \lambda\left(\sum\nolimits_{\omega\in\Omega} c_{od}^{i}(\omega)p_d(\omega) - I_d^i\right). \tag{C3}$$

The first order conditions imply

$$\frac{\partial\mathcal{L}}{\partial c_{od}^{i}(\omega)} = 0 \Leftrightarrow u_\omega = \lambda p_d(\omega), \tag{C4}$$

$$\sum\nolimits_{\omega\in\Omega} c_{od}^{i}(\omega)p_d(\omega) = I_d^i. \tag{C5}$$

Considering another commodity ω' and using Equations (C1) and (C4), we derive

$$c_{od}^{i}(\omega') = \left[\frac{p_d(\omega)}{p_d(\omega')}\right]^{\eta} c_{od}^{i}(\omega). \tag{C6}$$

Substituting Equation (C6) into Equation (C5) yields the optimal consumption on commodity ω:

$$c_{od}^{i}{}^{*}(\omega) = \frac{p_d(\omega)^{-\eta}}{P_d^{1-\eta}} I_d^i, \tag{C7}$$

where $P_d \equiv [\sum_{\omega\in\Omega} p_d(\omega)^{1-\eta}]^{1/(1-\eta)}$ is the aggregated price index in location d. Subsequently, substitute Equation (C7) into the utility function u to yield the maximal utility:

$$u^* = \frac{I_d^i}{P_d}. \tag{C8}$$

Finally, taking into account the migration costs μ_{od} and idiosyncratic component ϵ_{od}^i, and using Equation (3) along with $I_d^i = z_d^i \bar{I}_d$ and $v_{od}^i = \epsilon_{od}^i z_d^i$, we derive the indirect utility function as

$$V_{od}^i = \frac{v_{od}^i \bar{I}_d}{\bar{d}_{od} \tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od} P_d}, \tag{C9}$$

which is Equation (5). ■

C.2 Proof of Equation (7)

As demonstrated in Equation (6), the probability that individuals from location o choose to migrant to location d (or to remain locally if $d = o$) is the probability that the utility in location d exceeds that in any other locations. Specifically,

$$m_{od}^i = \Pr\left\{\frac{v_{od}^i \bar{I}_d}{\mu_{od} P_d} \geq \max_{d' \neq d}\left\{\frac{v_{od'}^i \bar{I}_{d'}}{\mu_{od'} P_{d'}}\right\}\right\}. \tag{C10}$$

To derive a closed-form expression, we have assumed that the idiosyncratic term v_{od}^i is independently and identically drawn from a Fréchet distribution, $F_v(x) = \mathrm{e}^{-x^{-\kappa}}$, with $x > 0$. Using the Law of Total Probability, we have

$$\begin{aligned} m_{od} &= \int_x \Pr\left\{\frac{x\bar{I}_d}{\mu_{od} P_d} \geq \max_{d' \neq d}\left\{\frac{v_{od'}^i \bar{I}_{d'}}{\mu_{od'} P_{d'}}\right\}\right\} \mathrm{d}F_v(x) \\ &= \int_x \Pr\left\{\frac{x\bar{I}_d}{\mu_{od} P_d} \geq \frac{v_{o1}^i \bar{I}_1}{\mu_{o1} P_1}, \cdots, \frac{x\bar{I}_d}{\mu_{od} P_d} \geq \frac{v_{oN}^i \bar{I}_N}{\mu_{oN} P_N}\right\} \mathrm{d}F_v(x). \end{aligned} \tag{C11}$$

Since $\{v_{od}^i\}_{d \in \mathcal{N}}$ are independent, we further derive

$$\begin{aligned} m_{od} &= \int_x \left(\prod\nolimits_{d' \neq d} \Pr\left\{\frac{x\bar{I}_d}{\mu_{od} P_d} \geq \frac{v_{od'}^i \bar{I}_{d'}}{\mu_{od'} P_{d'}}\right\}\right) \mathrm{d}F_v(x) \\ &= \int_x \left(\prod\nolimits_{d' \neq d} \mathrm{e}^{-\left[\frac{\bar{I}_d/(\mu_{od} P_d)}{\bar{I}_{d'}/(\mu_{od'} P_{d'})} x\right]^{-\kappa}}\right) \mathrm{d}F_v(x) \\ &= \int_x \mathrm{e}^{-\frac{\sum_{d' \neq d}[\bar{I}_{d'}/(\mu_{od'} P_{d'})]^{\kappa}}{[\bar{I}_d/(\mu_{od} P_d)]^{\kappa}} x^{-\kappa}} \mathrm{d}F_v(x). \end{aligned} \tag{C12}$$

Substitute the probability density, $\mathrm{d}F_v(x) = f_v(x)\mathrm{d}x = -\mathrm{e}^{-x^{-\kappa}}\mathrm{d}(x^{-\kappa})$, to yield

$$\begin{aligned} m_{od} &= -\int_x \mathrm{e}^{-\frac{\sum_{d' \neq d}[\bar{I}_{d'}/(\mu_{od'} P_{d'})]^{\kappa}}{[\bar{I}_d/(\mu_{od} P_d)]^{\kappa}} x^{-\kappa}} \mathrm{e}^{-x^{-\kappa}} \mathrm{d}(x^{-\kappa}) \\ &= -\int_x \mathrm{e}^{-\frac{\sum_{d' \in \mathcal{N}}[\bar{I}_{d'}/(\mu_{od'} P_{d'})]^{\kappa}}{[\bar{I}_d/(\mu_{od} P_d)]^{\kappa}} x^{-\kappa}} \mathrm{d}(x^{-\kappa}) \\ &= \frac{[\bar{I}_d/(\mu_{od} P_d)]^{\kappa}}{\sum_{d' \in \mathcal{N}}[\bar{I}_{d'}/(\mu_{od'} P_{d'})]^{\kappa}}. \end{aligned} \tag{C13}$$

Finally, substituting Equation (3) into Equation (C13), we derive

$$m_{od} = \frac{\bar{I}_d^{\kappa} P_d^{-\kappa} \left(\bar{d}_{od} \tilde{d}_{od} b_{od}^{p} b_{od}^{x} \lambda_{od}\right)^{-\kappa}}{\sum_{d' \in \mathcal{N}} \bar{I}_{d'}^{\kappa} P_{d'}^{-\kappa} \left(\bar{d}_{od'} \tilde{d}_{od'} b_{od'}^{p} b_{od'}^{x} \lambda_{od'}\right)^{-\kappa}}, \tag{C14}$$

which is Equation (7). ■

C.3 Proof of Equation (8)

Based on Equation (7), we can easily derive Equation (8). Using the definition of conditional probability, we have

$$\begin{aligned} m_{od,-o}^{i} &= \Pr\left\{V_{od}^{i} \geq \max_{d' \neq d}\{V_{od'}^{i}\} \,|\, V_{oo}^{i} < \max_{d' \neq o}\{V_{od'}^{i}\}\right\} \\ &= \frac{\Pr\left\{V_{od}^{i} \geq \max_{d' \neq d}\{V_{od'}^{i}\}, V_{oo}^{i} < \max_{d' \neq o}\{V_{od'}^{i}\}\right\}}{\Pr\left\{V_{oo}^{i} < \max_{d' \neq o}\{V_{od'}^{i}\}\right\}} \\ &= \frac{\Pr\left\{V_{od}^{i} \geq \max_{d' \neq d}\{V_{od'}^{i}\}\right\}}{\Pr\left\{V_{oo}^{i} < \max_{d' \neq o}\{V_{od'}^{i}\}\right\}} \\ &= \frac{\Pr\left\{V_{od}^{i} \geq \max_{d' \neq d}\{V_{od'}^{i}\}\right\}}{1 - \Pr\left\{V_{oo}^{i} \geq \max_{d' \neq o}\{V_{od'}^{i}\}\right\}}. \end{aligned} \tag{C15}$$

Since Equation (7) has provided the expressions for the two probabilities in Equation (C15), substituting them yields

$$m_{od,-o} = \frac{\bar{I}_d^{\kappa} P_d^{-\kappa} \left(\bar{d}_{od} \tilde{d}_{od} b_{od}^{p} b_{od}^{x} \lambda_{od}\right)^{-\kappa}}{\sum_{d' \neq o} \bar{I}_{d'}^{\kappa} P_{d'}^{-\kappa} \left(\bar{d}_{od'} \tilde{d}_{od'} b_{od'}^{p} b_{od'}^{x} \lambda_{od'}\right)^{-\kappa}}, \tag{C16}$$

which is Equation (8). ■

C.4 Imperfect information as a type of migration costs

In the main text, we treat information incompleteness as a form of migration cost, integrating it with transportation expenses modeled as a function of distance. This simplification avoids the need to solve the expected utility maximization problem that incorporates individuals' beliefs about destination income. We show that, under specific assumptions regarding individuals' beliefs and the relationship between distance and information incompleteness, the influence of imperfect income information on migration decision can be equivalently represented as a discount factor applied to utility, which provides a plausible rationale for this simplification.

Consider the scenario where income information for each destination d is imperfect; is imperfect; specifically, the realization of $\bar{I}_d$ is unknown ex-ante. However, individuals originating from location o share a common belief about its distribution, characterized by an expected value $\mathbb{E}[\bar{I}_d]$ and a variance $\mathrm{Var}_o[\bar{I}_d]$. We assume that $\mathbb{E}[\bar{I}_d]$ remains constant across origins but allow $\mathrm{Var}_o[\bar{I}_d]$ to vary by origin, capturing the greater information incompleteness associated with longer distance.

Suppose that individuals have a Constant Relative Risk Aversion (CRRA) utility function as follows:

$$U_{od}^{i} = \ln \frac{\epsilon_{od}^{i}}{\tilde{\mu}_{od}} C_{od}^{i} \,. \tag{C17}$$

Here $\tilde{\mu}_{od}$ denotes all bilaterial resistance factors defined in the main text excluding the component due to imperfect information. Since ϵ_{od}^{i}, $\tilde{\mu}_{od}$, and income deviation z_{d}^{i} attributed to individual characteristics are known by individuals prior to decision making, the sole source of uncertainly arises from the imperfect information regarding average income, specifically $\bar{I}_{d}$.

Following a reasoning similar to the proof of Equation (5), and given the nominal income $I_{d}^{i} = z_{d}^{i}\bar{I}_{d}$, the ex-post consumption decisions yield the following maximized utility for individuals:

$$V_{od}^{i}|\bar{I}_{d} = \ln \frac{\epsilon_{od}^{i}}{\tilde{\mu}_{od}} \frac{I_{od}^{i}}{P_{d}} = \ln \frac{\epsilon_{od}^{i} z_{od}^{i}}{\tilde{\mu}_{od} P_{d}} \bar{I}_{d} \,. \tag{C18}$$

Therefore, the expected value of the maximized utility is:

$$\mathbb{E}\left[V_{od}^{i}\right] = \mathbb{E}\left[\mathbb{E}\left[V_{od}^{i}|\bar{I}_{d}\right]\right] = \ln \frac{\epsilon_{od}^{i} z_{od}^{i}}{\tilde{\mu}_{od} P_{d}} + \mathbb{E}[\ln \bar{I}_{d}]. \tag{C19}$$

To aid comprehension, we begin with a simplified case where the impact of imperfect information can be incorporated as a utility discount factor, similar to the modeling of transportation expenses and culture-related costs discussed in the main text. Supposing that $\bar{I}_{d}$ is drawn from an exponential distribution with parameter λ_{d}; that is $f_{\bar{I}}(x) = \lambda_{d} \mathrm{e}^{-\lambda_{d} x}$, with $x \geq 0$, it is easy to prove that

$$\mathbb{E}[\bar{I}_{d}] = \frac{1}{\lambda_{d}}, \mathrm{Var}_{o}[\bar{I}_{d}] = \frac{1}{\lambda_{d}^{2}}, \mathbb{E}[\ln \bar{I}_{d}] = -\ln \lambda_{d} - \gamma,$$

where $\gamma \equiv -\int_{\mathbb{R}^{+}} \mathrm{e}^{-x} \ln x \, \mathrm{d}x$ is the Euler-Mascheroni constant. Therefore, we have

$$\mathbb{E}[\ln \bar{I}_{d}] = \ln(\mathbb{E}[\bar{I}_{d}]) - \ln \mathrm{e}^{\gamma} \,. \tag{C20}$$

Substituting Equation (C20) into Equation (C19), we derive

$$\mathbb{E}\left[V_{od}^{i}\right] = \ln \frac{\epsilon_{od}^{i} z_{od}^{i} \mathbb{E}[\bar{I}_{d}]}{\tilde{\mu}_{od} P_{d} \mathrm{e}^{\gamma}}. \tag{C21}$$

Here, $\mathrm{e}^{\gamma} > 1$ represents the utility discount factor. Accordingly, migration choices can be interpreted as individuals selecting their destination based on a shared belief in the expected income of location d, adjusted downward to account for uncertainty. However, a notable limitation is that this discount factor is constant and does not account for the varying influence of migration distance.

A more complex case assumes that $\bar{I}_{d}$ is drawn from a log-normal distribution, specifically $\ln \bar{I}_{d} \sim N_{o}(\theta_{od}, \sigma_{od}^{2})$ for individuals originating from location o. Under this assumption, if σ_{od}^{2} depends on physical distance as follows:

$$\sigma_{od}^{2} = \alpha + \beta \ln Distance_{od} \,, \tag{C22}$$

which reflects the idea that longer distances are associated with greater uncertainties, along with a relationship between θ_{od} and σ_{od}^{2},

$$\theta_{od} = \Theta_{d} - \frac{\sigma_{od}^{2}}{2}, \tag{C23}$$

where Θ_{d} is destination-specific, it can be shown that

$$\mathbb{E}[\bar{I}_{d}] = \mathrm{e}^{\theta_{od} + \frac{\sigma_{od}^{2}}{2}} = \mathrm{e}^{\Theta_{d}}, \tag{C24}$$

$$\mathrm{Var}[\bar{I}_d] = \left(\mathrm{e}^{\sigma_{od}^2} - 1\right)\mathrm{e}^{2\theta_{od}+\sigma_{od}^2} = \mathrm{e}^{2\Theta_d}\left(\mathrm{e}^{\alpha+\beta\ln Distance_{od}} - 1\right). \tag{C25}$$

This approach allows us to construct a set of distributions characterized by a shared belief in $\mathbb{E}[\bar{I}_d]$, while uncertainties, represented by $\mathrm{Var}[\bar{I}_d]$, systematically increase with distance.

Further, with Equation (C24) and $\mathbb{E}[\ln \bar{I}_d] = \theta_{od}$ we have

$$\mathbb{E}[\ln \bar{I}_d] = \ln(\mathbb{E}[\bar{I}_d]) - \frac{\sigma_{od}^2}{2} = \ln\left(\frac{\mathbb{E}[\bar{I}_d]}{\mathrm{e}^{\sigma_{od}^2/2}}\right). \tag{C26}$$

Substitute Equation (C22) into Equation (C26) to yield

$$\mathbb{E}[\ln \bar{I}_d] = \ln\frac{\mathbb{E}[\bar{I}_d]}{\mathrm{e}^{(\alpha+\beta\ln Distance_{od})/2}}, \tag{C27}$$

where the expected utility derived from random consumption is again expressed as the expected income discounted by a factor that accounts for income uncertainty, which grows with increasing migration distance. Substituting it into Equation (C27) yields

$$\mathbb{E}\left[V_{od}^i\right] = \ln\frac{\epsilon_{od}^i z_{od}^i \mathbb{E}[\bar{I}_d]}{\tilde{\mu}_{od} P_d \mathrm{e}^{(\alpha+\beta\ln Distance_{od})/2}}. \tag{C28}$$

To derive the same expression as presented in the main text, we decompose distance-related costs into two components: transportation expenses and income information incompleteness. Formally, $\bar{d}_{od} = \bar{d}_{od,trans}\bar{d}_{od,info}$, where the first term on the right-hand side represents transportation expenses, and the second term captures costs attributed to information incompleteness. Accordingly, we have

$$\tilde{\mu}_{od} = \bar{d}_{od,trans}\tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od}, \tag{C29}$$

and

$$\bar{d}_{od,info} = \mathrm{e}^{\frac{(\alpha+\beta\ln Distance_{od})}{2}}. \tag{C30}$$

Consistent with Equation (9), we model transportation expenses as following a constant elasticity:

$$\bar{d}_{od,trans}^{-\kappa} = \bar{\mu}_{trans} \cdot Distance_{od}^{\sigma_{trans}}. \tag{C31}$$

Define $\bar{\mu}_{info} \equiv \mathrm{e}^{-\kappa\alpha/2}$ and $\sigma_{info} \equiv -\kappa\beta/2$, we derive

$$\bar{d}_{od,info}^{-\kappa} = \bar{\mu}_{info} \cdot Distance_{od}^{\sigma_{info}}. \tag{C32}$$

Combining Equations (C31) and (C32), the total distance-related costs are

$$\bar{d}_{od}^{-\kappa} = \bar{\mu} \cdot Distance_{od}^{\sigma}, \tag{C33}$$

where $\bar{\mu} \equiv \bar{\mu}_{trans}\bar{\mu}_{info}$ and $\sigma \equiv \sigma_{trans}\sigma_{info}$, and this is Equation (9).

By incorporating the above modeling of distance-related costs that account for information uncertainty, individual's migration decision is expressed as

$$\begin{aligned} m_{od}^i &= \Pr\left\{\mathbb{E}\left[V_{od}^i\right] \geq \max_{d'\neq d}\left\{\mathbb{E}\left[V_{od'}^i\right]\right\}\right\} \\ &= \Pr\left\{\ln\frac{v_{od}^i \mathbb{E}[\bar{I}_d]}{\bar{d}_{od}\tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od} P_d} \geq \max_{d'\neq d}\left\{\ln\frac{v_{od'}^i \mathbb{E}[\bar{I}_{d'}]}{\bar{d}_{od'}\tilde{d}_{od'} b_{od'}^p b_{od'}^x \lambda_{od'} P_{d'}}\right\}\right\} \\ &= \Pr\left\{\frac{v_{od}^i \mathbb{E}[\bar{I}_d]}{\bar{d}_{od}\tilde{d}_{od} b_{od}^p b_{od}^x \lambda_{od} P_d} \geq \max_{d'\neq d}\left\{\frac{v_{od'}^i \mathbb{E}[\bar{I}_{d'}]}{\bar{d}_{od'}\tilde{d}_{od'} b_{od'}^p b_{od'}^x \lambda_{od'} P_{d'}}\right\}\right\}, \end{aligned} \tag{C34}$$

which mirrors the expression in the main text, except that the known incomes are replaced by a common

belief in their expected values.

D Estimating the gravity equation

This appendix discusses our estimation strategy. First, we compare two commonly used estimation methods for multiplicative models or gravity equations: Poisson Pseudo Maximum Likelihood (PPML) and Ordinary Least Squares (OLS) with log or log-like transformation. Second, we explain how PPML estimates are obtained and discuss their relationship with the theoretical model presented in the main text. Specifically, we show that the log-likelihood function maximized by the PPML estimator is identical to the one derived from our theoretical model; in other words, PPML serves as a convenient tool for estimating our model. Finally, we briefly prove the consistency of the PPML estimator and use Monte Carlo simulations to verify that, given the sample size of our available migrant records, PPML performs well. With the same simulated sample, we compare PPML with OLS using log-linearization or log-like transformations, finding that both OLS methods yield significantly biased estimates.

In summary, our discussion of estimation methods offers a useful suggestion for applied studies involving multiplicative models, such as those in trade or migration, particularly in economic history research where data are often scarce. We observe that, even when the sample size of individuals is much smaller than the number of location pairs, and most pairs exhibit zero migrant flow, PPML provides reliable estimates for model parameters, given the model specification is correct. In contrast, OLS with transformations of the dependent variable performs poorly. Therefore, researchers may consider basing their interpretations on estimates provided by PPML, particularly when the sample size is small.

D.1 Comparison between OLS and PPML

Multiplicative models are widely utilized in the economics literature, particularly in studies on trade and migration, where gravity equations are often employed to describe bilateral trade flows or labor mobility. A typical gravity equation predicts bilateral economic activities as follows:

$$Y_{od} = \alpha_0 X_o X_d D_{od}, \tag{D1}$$

where Y_{od} denotes the volume of trade or migration between two regions, X_o and X_d capture the specific characteristics of the origin and destination, respectively, and D_{od} represents the bilateral resistance. To address deviations from theoretical predictions, empirical studies often adopt stochastic versions of the gravity equation. Typically,

$$\begin{gathered} Y_{od} = \alpha_0 X_o^{\alpha_1} X_d^{\alpha_2} D_{od}^{\alpha_3} \eta_{od}, \\ \text{or equivalently, } Y_{od} = \exp(\ln \alpha_0 + \alpha_1 \ln X_o + \alpha_2 \ln X_d + \alpha_3 \ln D_{od}) + \varepsilon_{od}. \end{gathered} \tag{D2}$$

where η_{od} is an error term being mean-independent of the other regressors. Formally, it satisfies $\mathbb{E}[\eta_{od}|X_o, X_d, D_{od}] = \mathbb{E}[\eta_{od}] = 1$. In the equivalent expression, $\varepsilon_{od} \equiv \alpha_0 X_o^{\alpha_1} X_d^{\alpha_2} D_{od}^{\alpha_3}(\eta_{od} - 1)$, which satisfies $\mathbb{E}[\varepsilon_{od}|X_o, X_d, D_{od}] = \alpha_0 X_o^{\alpha_1} X_d^{\alpha_2} D_{od}^{\alpha_3} \mathbb{E}[\eta_{od} - 1|X_o, X_d, D_{od}] = 0$. Therefore, we have

$$\mathbb{E}[Y_{od}|X_o, X_d, D_{od}] = \alpha_0 X_o^{\alpha_1} X_d^{\alpha_2} D_{od}^{\alpha_3}. \tag{D3}$$

As discussed in the main text, two common methods are used to estimate Equation (D2) in applied studies. The first involves taking the logarithm of both sides, resulting in a linear specification,

$$\ln Y_{od} = \ln \alpha_0 + \alpha_1 \ln X_o + \alpha_2 \ln X_d + \alpha_3 \ln D_{od} + \ln \eta_{od}\,, \tag{D4}$$

which can then be estimated using OLS. However, as noted by Silva and Tenreyro (2006), this log-linearization approach has at least two notable limitations. The first limitation is that obtaining unbiased estimates using OLS requires $\mathbb{E}[\ln \eta_{od}\,|X_o, X_d, D_{od}] = 0$, however, $\mathbb{E}[\eta_{od}|X_o, X_d, D_{od}] = 1$ does not

necessarily imply $\mathbb{E}[\ln \eta_{od} | X_o, X_d, D_{od}] = 0$. This is because the latter depends not only on the conditional mean of η_{od} but also on the higher-order moments of its conditional distribution. For instance, in the presence of heteroskedasticity—where $\mathrm{Var}[\eta_{od}|X_o, X_d, D_{od}]$ is a function of X_o, X_d, and D_{od}—$\ln \eta_{od}$ is unlikely to be mean-independent of these variables. This issue arises because the nonlinear transformation of the dependent variable "changes the properties of the error term in a nontrivial way."

The second limitation pertains to samples where $Y_{od} = 0$. While such observations pose no issue for estimating gravity equations in their multiplicative form, they must be excluded in the log-linearized form, as the logarithmic function is undefined at zero. A commonly adopted workaround is to apply log-like transformations, such as $\ln(Y_{od} + 1)$ or $\ln \sqrt{Y_{od} + (1 + Y_{od}^2)}$. However, some recent studies highlight that these transformations can introduce significant estimation biases, particularly when extensive-margin effects—that is, the impact of zero versus nonzero observations—are substantial and cannot be ignored. With log-like transformations, the estimated coefficients "can be made to take any desired value through the appropriate choice of [the units of the dependent variable]" (Chen and Roth 2024).

Another estimation method is to use the PPML estimator. For better understanding, we focus on our estimation equation from the main text, Equation (12), and rewrite it in the following expectation form under the assumption $\mathbb{E}[\varepsilon_{od}|X_{od}] = 0$:

$$\mathbb{E}\big[m_{od,-o}|X_{od}\big] = \frac{\varphi_{od}}{\sum_{d' \neq o} \varphi_{od'}}, \tag{D5}$$

where

$$\begin{aligned} \varphi_{od} &\equiv \exp(X_{od}'\beta) \\ &\equiv \exp[\,\gamma + \beta_d + \sigma \ln Distance_{od} + \pi(1 - Culture_{od}) \\ &\quad + \big(\ln \bar{b}^p\big)\mathbb{I}\{Prov_o \neq Prov_d\} + \big(\ln \bar{b}^x\big)\mathbb{I}\{Xiaqu_o \neq Xiaqu_d\}]. \end{aligned} \tag{D6}$$

It should be noted that α_o is not included in φ_{od}, as it appears in the denominator of Equation (D5).

We begin by introducing an assumption, which is ultimately unnecessary, that the observed number of migrants originating from location o and migrating to location d, denoted by l_{od}, is drawn from a Poisson distribution with mean $\delta_o \varphi_{od}$, where δ_o denotes the origin fixed effects. Therefore, the probability of observing l_{od} is given by

$$f(l_{od}; \beta, \delta_o) = \frac{(\delta_o \varphi_{od})^{l_{od}} \exp(-\delta_o \varphi_{od})}{l_{od}!}. \tag{D7}$$

Therefore, the log-likelihood function for the complete sample is

$$\mathcal{L}\big(\{l_{od}\}_{o,d}; \beta, \{\delta_o\}_o\big) = \sum\nolimits_o \sum\nolimits_{d \neq o} l_{od} \ln \delta_o + \sum\nolimits_o \sum\nolimits_{d \neq o} l_{od} \ln \varphi_{od} - \sum\nolimits_o \sum\nolimits_{d \neq o} \delta_o \varphi_{od}\,, \tag{D8}$$

where we ignore a constant term.

To maximize the above log-likelihood, the first order condition with respect to the destination fixed effects implies

$$\hat{\delta}_o = \frac{\sum_{d \neq o} l_{od}}{\sum_{d \neq o} \varphi_{od}}. \tag{D9}$$

Substituting Equation (D9) into (D8) yields

$$\mathcal{L}\big(\{l_{od}\}_{o,d}; \beta\big) = \sum\nolimits_o \sum\nolimits_{d \neq o} l_{od} \ln\left(\frac{\varphi_{od}}{\sum_{d \neq o} \varphi_{od}}\right), \tag{D10}$$

where a constant term is ignored again. The first order condition with respect to $\hat{\beta}$ is

$$\sum\nolimits_o l_o \sum\nolimits_{d\neq o} \left(\frac{l_{od}}{l_o} - \frac{\hat{\varphi}_{od}}{\sum_{d'\neq o}\hat{\varphi}_{od'}}\right) X_{od} = 0, \tag{D11}$$

where $l_o \equiv \sum_{d\neq o} l_{od}$ and $\hat{\varphi}_{od} = \exp\left(X'_{od}\hat{\beta}\right)$. Solving the Equation (D11) yields the PPML estimator for β, encompassing all coefficients of interest.

D.2 The linkage between the PPML estimator and the structural gravity equation

Is it justified to assume that the observed data follows a Poisson distribution? Referring to Sotelo (2019), here we demonstrate that assuming a specific type of distribution is not strictly necessary; instead, PPML just serves as a tool to conveniently estimate.

Given that our discrete migration choice model suggests a specific data-generating process, a more natural approach is to interpret the observed migration choices as a finite sample from this process and maximize its likelihood. To recap, $m_{od,-o}$ is the conditional probability predicted by our theoretical model that an individual i originating from location o migrates to location d. Denoting the observed number of corresponding migrants as l_{od} and assuming that individuals are independent, the likelihood of the observed data is $\prod_{o,d\neq o}(m_{od,-o})^{l_{od}}$. Thus, the log-likelihood function is derived as follows:

$$\begin{aligned}\mathcal{L}'\left(\{l_{od}\}_{o,d};\beta\right) &\equiv \sum\nolimits_o \sum\nolimits_{d\neq o} l_{od} \ln m_{od,-o} \\ &= \sum\nolimits_o \sum\nolimits_{d\neq o} l_{od} \ln\left(\frac{\varphi_{od}}{\sum_{d\neq o}\varphi_{od}}\right),\end{aligned} \tag{D12}$$

which, notably, is identical to the log-likelihood function in Equation (D10). As highlighted by Sotelo (2019), this numerical equivalence is particularly useful when the Poisson routine is readily available, whereas the routine for directly maximizing the above likelihood is not. To avoid the need for programming the latter, we can "trick" the Poisson estimator into performing the task by setting the conditional migrant share as the dependent variable and adding the origin fixed effects.

Revisiting the first-order condition in Equation (D11) provides a more intuitive understanding. In this context, $l_{od}/l_o \equiv m_{od,-o}$ is the observed migrant share, and $\hat{\varphi}_{od}/\sum_{d'\neq o}\hat{\varphi}_{od'} \equiv \hat{m}_{od,-o}$ is the predicted migrant share based on the theoretical model (see Equation (D5)). Consequently, the first-order condition ensures that the weighted sum of estimated residuals $m_{od,-o} - \hat{m}_{od,-o}$ equals zero for each explanatory variable in X_{od}.

D.3 The consistency of the PPML estimator and Monte Carlo simulation

Finally, we discuss the consistency of the PPML estimator. We first provide a brief proof and then supplement it with a Monte Carlo simulation based on our migration choice model in the main text. The property of consistency requires that, as long as the model is correctly specified (i.e., $\mathbb{E}[\varepsilon_{od}|X_{od}] = 0$ holds), maximizing Equation (D10) offers $\hat{\beta}^{\text{PPML}} \to_p \beta$.

By definition, given $\mathbb{E}[\varepsilon_{od}|X_{od}] = 0$, we have

$$\mathbb{E}\left[\frac{l_{od}}{l_o}|X_{od}\right] = \frac{\varphi_{od}}{\sum_{d'\neq o}\varphi_{od'}}, \tag{D13}$$

where $l_o \equiv \sum_{d\neq o} l_{od}$. Therefore, the expectation of the score function $S\left(\hat{\beta}\right) \equiv \partial\mathcal{L}\left(\hat{\beta}\right)/\partial\hat{\beta}$ is

$$\begin{aligned}\mathbb{E}\left[S(\hat{\beta})|X_{od}\right] &= \mathbb{E}\left[\sum\nolimits_o l_o \sum\nolimits_{d\neq o}\left(\frac{l_{od}}{l_o}-\frac{\hat{\varphi}_{od}}{\sum_{d'\neq o}\hat{\varphi}_{od'}}\right)X_{od}\,|X_{od}\right]\\ &= \sum\nolimits_o l_o \sum\nolimits_{d\neq o}\left(\mathbb{E}\left[\frac{l_{od}}{l_o}|X_{od}\right]-\frac{\hat{\varphi}_{od}}{\sum_{d'\neq o}\hat{\varphi}_{od'}}\right)X_{od}\,,\end{aligned} \tag{D14}$$

which takes the value of zero only for the true β. In other words, the true β is the maximizer of the expected log-likelihood, which satisfies

$$\mathbb{E}[S(\beta)] = 0. \tag{D15}$$

Using the Law of Large Number, as the sample size $n \to \infty$, the sample log-likelihood converges to the expected log-likelihood:

$$S(\hat{\beta}) \to_p \mathbb{E}[S(\hat{\beta})]. \tag{D16}$$

It is easy to prove that the log-likelihood function is concave in $\hat{\beta}$ due to the properties of the exponential family, thus the solution to $\mathbb{E}[S(\hat{\beta})] = 0$ is unique and corresponds to the global maximum at the true β. Therefore, when the sample is large enough, $S(\hat{\beta}) = 0$ implies that $\hat{\beta}$ must lie near the true β; that is, $\hat{\beta} \to_p \beta$.

Although the consistency of the PPML estimator is theoretically sound, it does not guarantee that the estimates will reflect the true underlying parameters in practice. For instance, if the sample size of migrants is insufficient relative to the number of location pairs, idiosyncratic preferences could significantly distort, or even bias, the estimations and calibrations, as highlighted by Dingel and Tintelnot (2020).

To address this issue, we simulate several migrant samples with the same size as our observed data. We then assign predetermined values to all model parameters (e.g., distance elasticity, province boundary effects, *xiaqu* boundary effects) and simulate migration decisions based on our theoretical model. Finally, we calculate the migrant share for each prefecture pair and use this simulated data to estimate the model parameters using the PPML estimator. By comparing the predetermined parameters with the estimated ones, we can assess the estimation accuracy of our data and methodology. The simulation procedure is as follows:

1. Extract all migrants recorded in the *XKTB* for the first period (i.e., the 1760s) and exclude those lacking information on their origin prefecture, resulting in a subsample of 4,177 effective migrants and 240 origin prefectures with recorded out-migrants. We restrict our simulation to these 240×240 prefecture pairs and assume that the 4,177 migrants originate from their respective recorded origin prefectures. The primary task is to construct several 4,177×240 matrices, each elements representing the different utility components (real income, migration costs, and the idiosyncratic component) that each migrant would derive from migrating to each destination prefecture.
2. Draw the real income for each prefecture from a log-normal distribution, $\ln \bar{I}_d \sim N(0,1)$, generating a 1×240 vector. This vector is then copied across to form a 4,177×240 income matrix $\mathbb{I} = [\bar{I}_d]_{4,177\times240}$, where each row represents the real income of all potential destination prefectures, with the same value across all rows.
3. Construct the migration distance matrix $[d_d^i]_{4,177\times240}$, where d_d^i denotes the physical migration distance for individual i migrating from her/his origin prefecture to the destination prefecture d. If prefecture d is the same as the origin prefecture of individual i, we set the distance to 999,999,999 kilometers to impose a sufficiently large migration cost, ensuring that all individuals decide to migrate. Assuming that the cost due to physical distance is $\bar{d}_{od}^{-1} = Distance_{od}^{-2}$, which corresponds to setting σ to -3 (=-2×κ, where κ is set to 1.5, as detailed below) and $\bar{\mu}$ to 1 in Equation (9), the resulting cost matrix for each individual's potential migration destination is

$$\mathbb{D} \equiv \left[\left(d_d^i\right)^{-2}\right]_{4,177\times240}.$$

4. Construct the clan culture difference matrix $[c_d^i]_{4,177\times240}$, where c_d^i represents one minus the cosine similarity of surname distributions between individual i's origin prefecture and destination prefecture d. Similarly, we set $\tilde{d}_{od} = \mathrm{e}^{1-Culture_{od}}$, which corresponds to setting π to -1.5 (=-1×κ) and $\tilde{\mu}$ to 1 in Equation (10). The clan culture cost matrix is then

$$\mathbb{C} \equiv \left[e^{c_d^i}\right]_{4,177\times240}.$$

5. Construct the matrix representing province-crossing journeys $[p_d^i]_{4,177\times240}$, where p_d^i takes the value of one if individual i's origin prefecture is located in a different province from destination prefecture d. Assuming that crossing province boundary won't lead to additional costs, which is consistent to our observations reported in the main text. This corresponds to setting $\ln \bar{b}^p$ to 0 in Equation (11). Then, the cost related to province boundary effects is

$$\mathbb{P} \equiv [1]_{4,177\times240}.$$

6. Construct the matrix representing province-crossing journeys $[x_d^i]_{4,177\times240}$, where x_d^i takes the value of one if individual i's origin prefecture is located in a different *xiaqu* from destination prefecture d. Assuming the utility discount factor of crossing *xiaqu* boundaries is 1.3, which corresponds to setting $\ln \bar{b}^p$ to -0.394 ($= -\kappa \cdot \ln 1.3$) in Equation (11), then the cost related to *xiaqu* boundary effects is

$$\mathbb{X} = \left[1.3^{-x_d^i}\right]_{4,177\times240}.$$

7. Draw 57,600 (=240×240) random utility discount factors $\{\lambda_{od}\}_{o,d}$, which represent the unobserved components, independently and identically drawn from a standard log-normal distribution. Add one to each draw to ensure all values are greater than one. Then, construct the unobserved disutility matrix $\mathbb{R} = [\lambda_d^i]_{4,177\times240}$, where λ_d^i denotes the corresponding draw for individual i's origin prefecture and destination prefecture d.
8. Draw 1,002,480 (=4,177×240) idiosyncratic utility components independently and identically from a Fréchet distribution, $F_v(x) = \mathrm{e}^{-x^{-\kappa}}$, with $x > 0$. The shape parameter κ is set to 1.5 based on the estimation by Tombe and Zhu (2019). Then, construct the idiosyncratic utility matrix $\mathbb{v} = [v_d^i]_{4,177\times240}$, where v_d^i denotes the idiosyncratic preference of individual i for migrating to destination prefecture d.
9. Taken together, the utility for each individual's potential destination $\mathbb{V}$ is given by the Hadamard product of the above matrices, i.e., $\mathbb{V} = \mathbb{v} \circ \mathbb{I} \circ \mathbb{D} \circ \mathbb{C} \circ \mathbb{P} \circ \mathbb{X} \circ \mathbb{R}$. The optimal migration decision for each individual is determined by identifying the maximum value in each row of this utility matrix.
10. Using these 4,177 simulated migrants, we construct the prefecture-pair dataset. We then estimate Equation (12) with the different estimators and compare the estimated coefficients to their predetermined values.

We conduct the simulation process described above 1,000 times, yielding 1,000 estimated coefficients for each regressor. Figure D.1 presents the estimated results with three estimators: the PPML estimator, the OLS estimator with log-linearization of the dependent variable, and the OLS estimator with a log-like transformation $\ln(m_{od,-o} + 0.01)$ applied to the dependent variable. The solid diamond markers represent the predetermined parameter values. Each hollow marker indicates the mean of the corresponding 1,000 estimated coefficients, while the intervals represent the range between the 5th and 95th percentiles.

Across the results, the point estimates from the PPML estimator are close to the true values and statistically significant for the non-zero parameters. For the province boundary effects, which we set to zero, the estimated coefficients are insignificant, rejecting the null hypothesis that province boundary effects exist. In contrast, the other two estimators show large deviations. This comparison yields two key implications. First, in the context of this study, the PPML estimator performs significantly better than the other estimators. More importantly, given the size of our available migrant sample, the PPML estimator can deliver credible estimates of administrative boundary effects that closely align with the true values, provided that our theoretical model is correctly specified.

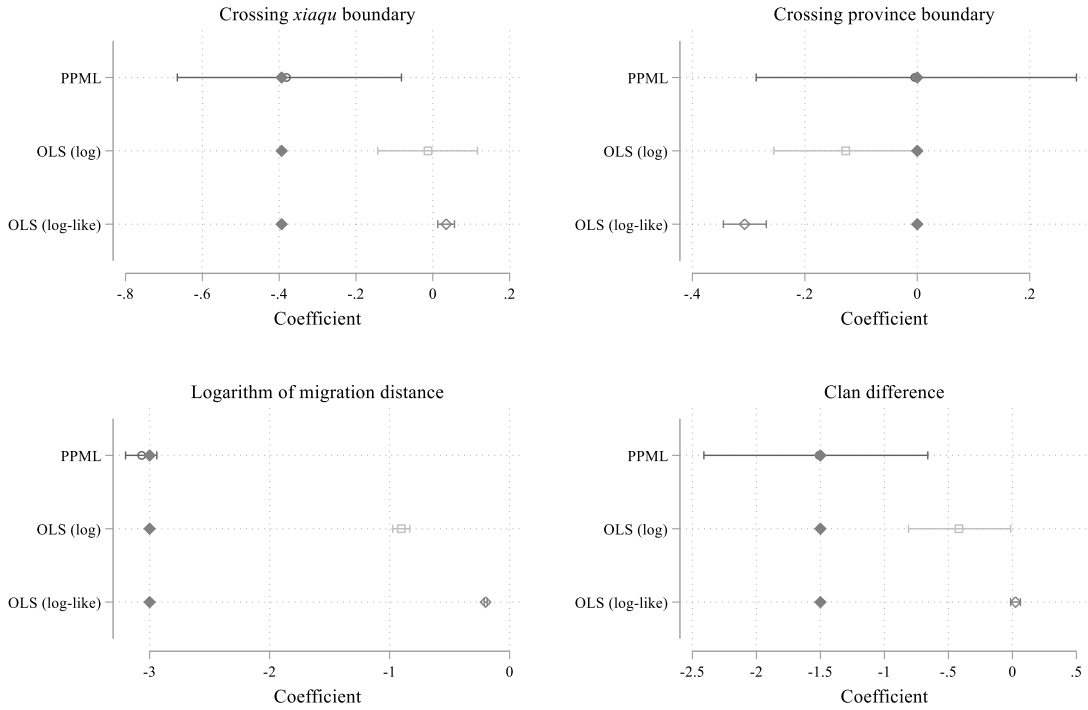

Figure D.1 The estimated coefficients derived from different estimators with 1,000 simulated samples.
Notes: This figure presents the estimated coefficients using PPML, OLS (with log-linearization, i.e., $\ln m_{od,-o}$ as the dependent variable), and OLS (with log-like transformation, i.e., $\ln(m_{od,-o}+0.01)$ as the dependent variable). In each figure, the solid diamond denotes the true value setting for producing simulated samples. Each hollow marker indicates the mean of the corresponding 1,000 estimated coefficients, while the intervals represent the range between the 5th and 95th percentiles.

In the main text, we use the estimates of destination fixed effects to recover the real income for each prefecture. To validate this procedure, we compare the estimated real incomes with those obtained through the simulation process. According to Equation (12), the destination fixed effects reflect $\kappa \ln(\bar{I}_d/P_d)$. However, since we can only identify $N-1$ fixed effects in the estimation, the estimated fixed effects are relative. Denoting the reference prefecture by the subscript r, the estimated destination fixed effect for prefecture d is

$$\hat{\beta}_d \to_p \kappa \ln(\bar{I}_d/P_d) - \kappa \ln(\bar{I}_r/P_r)\,. \tag{D17}$$

We center these fixed effects by subtracting their mean, expressed as:

$$\hat{\beta}_d^{\mathrm{c}} \equiv \hat{\beta}_d - \frac{1}{N}\sum\nolimits_{d'} \hat{\beta}_{d'} \to_p \kappa \ln\left[\frac{\bar{I}_d/P_d}{\prod_{d'}(\bar{I}_{d'}/P_{d'})^{1/N}}\right]. \tag{D18}$$

Accordingly, the relative real income of prefecture d is given by

$$\frac{\bar{I}_d/P_d}{\prod_{d'}(\bar{I}_{d'}/P_{d'})^{1/N}} = \exp\left(\frac{\hat{\beta}_d^{\mathrm{c}}}{\kappa}\right). \tag{D19}$$

Notably, the left-hand side of Equation (D19) can be constructed using the true values from the simulation, while the right-hand side is derived from a transformation of the estimated fixed effects. Comparing them enables us to evaluate the accuracy of the estimates of real incomes. For clarity, we present the results from one of the 1,000 simulations, generated with a random seed of 10000000, in Figure D.2. Each dot in the figure represents the true real income and the estimated real income for a specific prefecture. If the relative real income estimates are accurate, the points should align closely with the 45° line. While the limited migrant sample size introduces some fluctuation, the estimated results demonstrate a strong correlation with the true values.

Moreover, we find that the slope of the fitted line through these dots is close to, but slightly exceeds, 1. This pattern is observed in most of the 1,000 samples and indicates that the estimated fixed effects slightly overstate the relative real income. One possible explanation for this is that prefectures with lower income levels attract no migrants, due to the relatively small sample size, resulting in the model being unable to estimate fixed effects for these locations. This aligns with our findings, where estimations for each simulated sample consistently omit a small number of fixed effects. The remaining fixed effects, which are retained, may tend to be overestimated. However, since our main analysis focuses primarily on the changes in relative real income between pairs of prefectures, the impact of this minor systematic overestimation is likely to be mitigated.

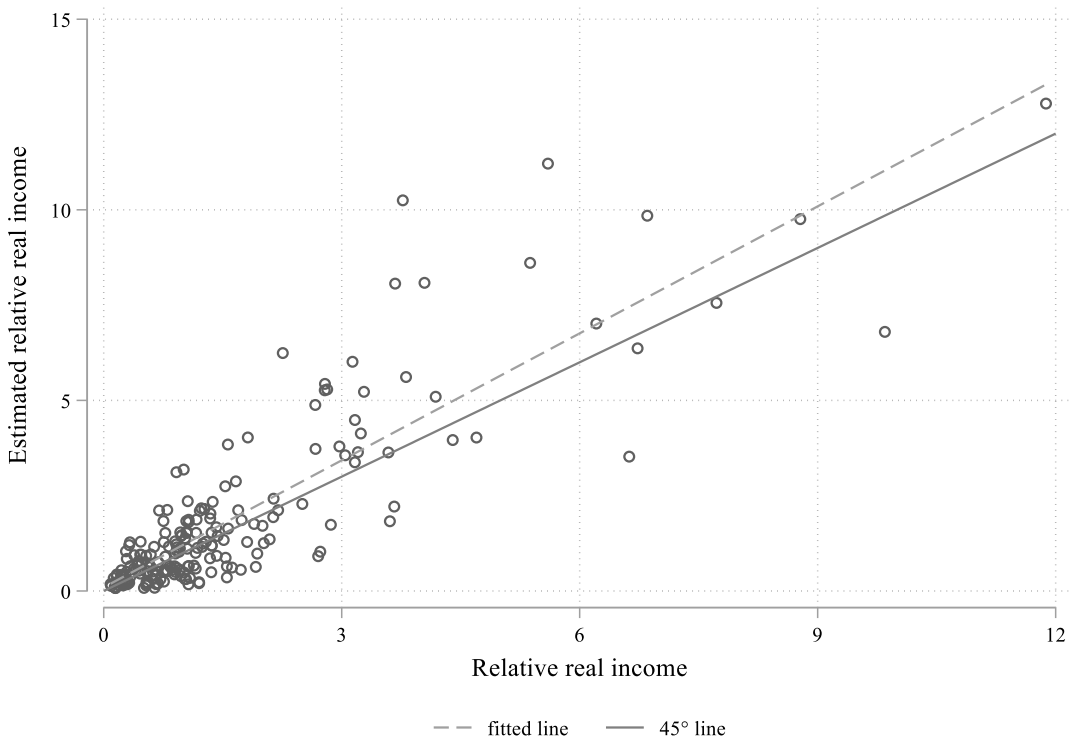

Figure D.2 Comparison between the estimated relative real incomes and the true values.
Notes: This figure compares the estimated relative real incomes, derived from destination fixed effects, with the true values. These estimates are yielded from a single simulation with the random seed 10000000. Each dot represents a prefecture, the dashed line denotes the fitted line, and the solid line denotes the 45° line.

To provide a general assessment, we conduct the comparison for each of the 1,000 simulated samples and

record the corresponding fitted R-squared values. Figure D.3 presents the kernel density of these R-squared values, with a mean of 0.695 and a standard deviation of 0.093.

We also estimate provincial-level average real income in the main text to compare our estimates with existing ones and to examine the impact of *xiaqu* boundaries on income fluctuations. This data structure, which combines both local residents and migrants, results in a migrant sample size that far exceeds the number of province pairs. However, this raises another concern: the small number of observations for each province in the pair data may lead to inconsistencies in the fixed effects, a common issue in short-panel data. To examine this, we use a similar procedure to construct 1,000 simulated province-pair datasets and compare the estimated provincial relative real incomes with those from the simulations. The kernel density of the resulting 1,000 R-squared values is shown in Figure D.4. We observe that the peak of the distribution is notably shifted to the right, and the degree of volatility has also increased.

To summarize, the fixed-effects estimates with the prefecture-pair data may be less accurate than those with the province-pair data partially due to the migrant sample size. However, due to the large number of observations across prefectures, the patterns observed in the different simulations tend to be more consistent. On the other hand, the fixed effects estimates for the province-pair data address the issue of smaller sample sizes relative to the number of location pairs. However, this approach introduces challenges similar to those encountered in short-panel data. While, on average, the estimates for the provincial data are significantly more accurate, the fluctuations between simulations become more pronounced.

Overall, these results suggest, with optimism, that, assuming the correct model specification, the destination fixed effects estimated from our migration data should generally reflect the relative real income level of each location.

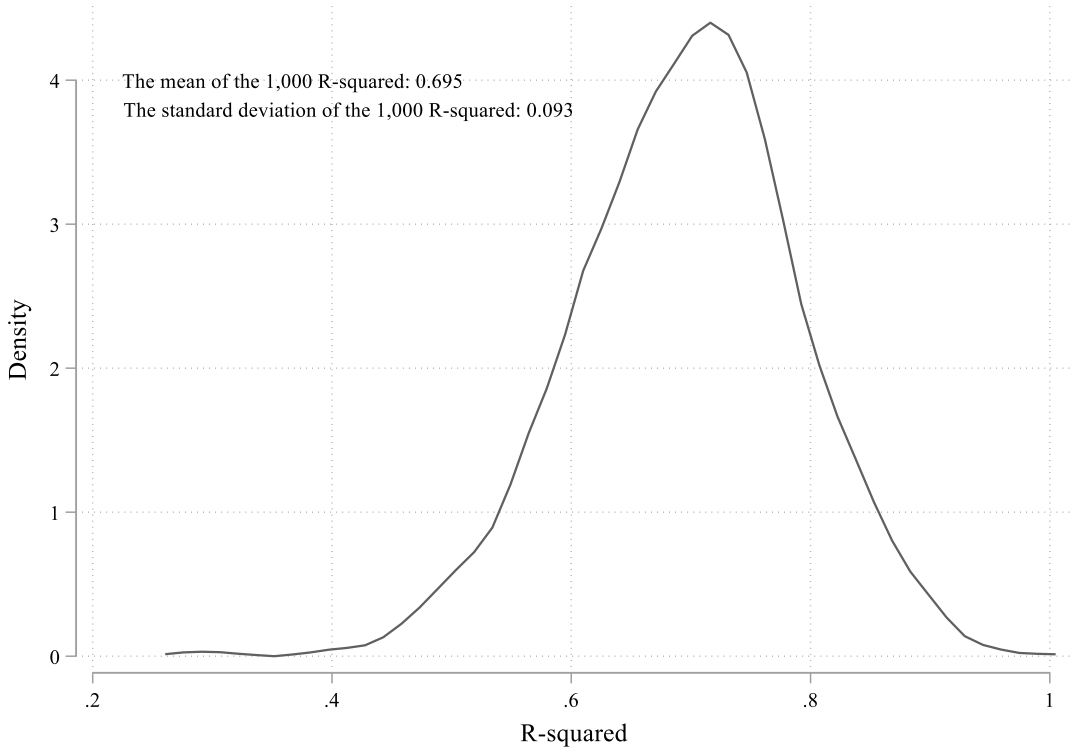

Figure D.3 Kernel density of the fitted R-squared between true and estimated real income, with prefecture-pair data.
Notes: This figure presents the kernel density of the fitted R-squared values between the true relative real incomes and those estimated from the destination fixed effects, based on 1,000 simulated prefecture-pair datasets.

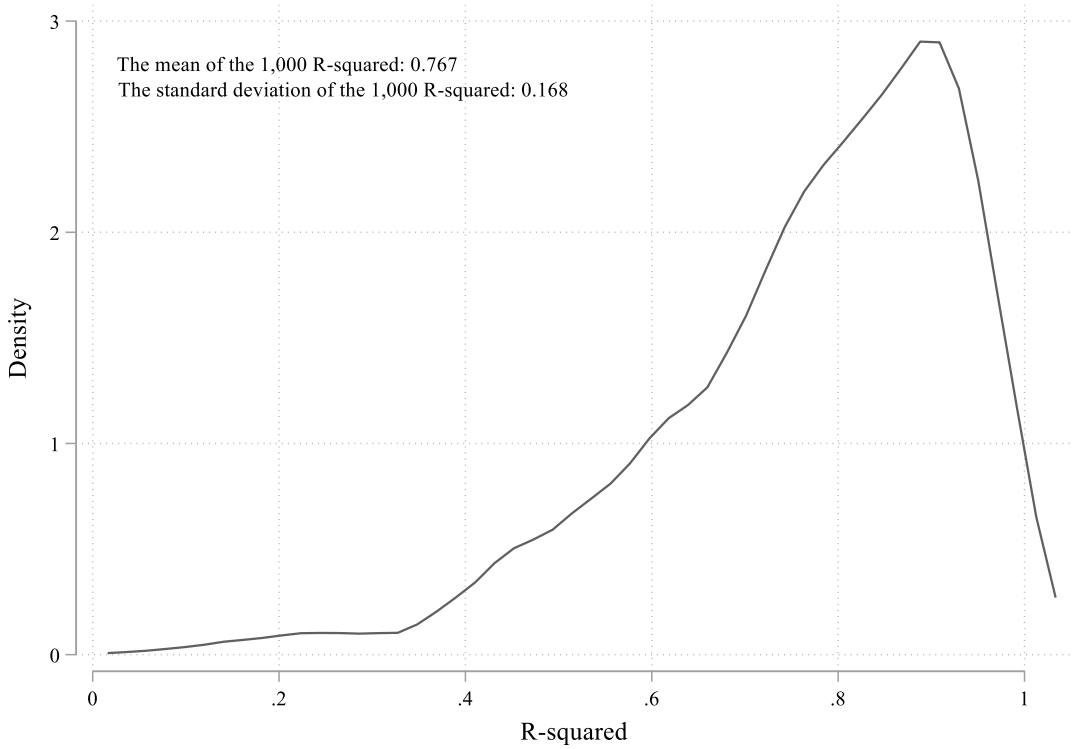

Figure D.4 Kernel density of the fitted R-squared between true and estimated real income, with province-pair data.
Notes: This figure presents the kernel density of the fitted R-squared values between the true relative real incomes and those estimated from the destination fixed effects, based on 1,000 simulated province-pair datasets.

E Estimating the shape parameter

This appendix provides estimates for the shape parameter κ in the Fréchet distribution, from which idiosyncratic preferences are drawn. The value of κ is important for two reasons. First, it allows us to establish a quantitative relationship between the estimated coefficients and the model parameters. For instance, the calculation of the utility discount factor $b_{od}^{x} = \mathrm{e}^{-\ln \bar{b}^{x}/\kappa}$ requires the value of κ. Second, knowing κ is essential if we aim to recover the relative real income of each prefecture or province, as the estimated destination fixed effect is a function of both relative real income and κ.

To begin with, suppose we have data on the average real income $\bar{I}_d/P_d$. In this case, Equation (7), in conjunction with the modeling of specific migration costs, provides a specification for estimating κ:

$$\begin{aligned} m_{od} = \exp[\, & \gamma + \alpha_o + \kappa(\bar{I}_d/P_d) + \sigma \ln Distance_{od} + \pi(1 - Culture_{od}) \\ & + \left(\ln \bar{b}^{p}\right)\mathbb{I}\{Prov_o \neq Prov_d\} + \left(\ln \bar{b}^{x}\right)\mathbb{I}\{Xiaqu_o \neq Xiaqu_d\}] + \varepsilon_{od}. \end{aligned} \tag{E1}$$

Assuming that ε_{od} is uncorrelated with real income, regressing the migrant shares on real income, along with other regressors and origin fixed effects, yields an estimate of κ.

Based on Equation (7), we derive

$$\frac{m_{od}}{m_{oo}} = \frac{\bar{I}_d^{\kappa} P_d^{-\kappa}}{\bar{I}_o^{\kappa} P_o^{-\kappa}} \left(\bar{d}_{od} \tilde{d}_{od} b_{od}^{p} b_{od}^{x} \lambda_{od}\right)^{-\kappa}, \tag{E2}$$

from which we derive an alternative specification for estimating κ:

$$\begin{aligned} \frac{m_{od}}{m_{oo}} = \exp[\, & \gamma + \kappa \left(\frac{\bar{I}_d/P_d}{\bar{I}_o/P_o}\right) + \sigma \ln Distance_{od} + \pi(1 - Culture_{od}) \\ & + \left(\ln \bar{b}^{p}\right)\mathbb{I}\{Prov_o \neq Prov_d\} + \left(\ln \bar{b}^{x}\right)\mathbb{I}\{Xiaqu_o \neq Xiaqu_d\}] + \varepsilon_{od}. \end{aligned} \tag{E3}$$

In this specification, we regress the ratio of migrant share to local share on relative real income, along with other regressors, without incorporating the origin fixed effects. It is important to note that these two specifications imply different orthogonal assumptions about the error terms. Therefore, we estimate both of them.

However, the key challenge is that the income levels are unknown to us and are, in fact, what we aim to recover. Consequently, we must rely on existing wage estimates to estimate κ. The primary source for this is Liu (2024), who provides estimates of provincial relative wage levels during 1530 and 1640. Liu (2024) collects 1,017 quotations of remuneration for *yi*/*ya yi* during this period. With this dataset, the author performs a regression of the logarithm of these 1,017 recorded wages on a set of region fixed effects, controlling for administrative levels, job types, and payment methods. The region fixed effects thus capture the relative average wage level for each region.

In Table A3 of Liu (2024), the author reports the estimated values for 11 regions, including Beijing, North Zhili, Shandong, Henan, Zhejiang, Jiangsu, Anhui, Jiangxi, Huguang (Hunan & Hubei), Guangdong, and Fujian. Eight of these regions can be directly matched to our province-pair dataset. We present the author's estimates in Table E.1. In Liu's estimation, Zhejiang serves as the reference region, so each estimate reflects the relative wage level. For instance, the value of -0.081 for Jiangsu's fixed effect indicates that the average wage in Jiangsu is approximately 91.9% of the average wage in Zhejiang.

Let $\{I_d\}_d$ represent the nominal wage levels and $\{\hat{\gamma}_d\}_d$ denote the estimated fixed effects. We can express their relationship as follows:

$$\hat{\gamma}_d \to_p \ln I_d - \ln I_{\text{Zhejiang}} = \ln \frac{I_d}{I_{\text{Zhejiang}}}. \tag{E4}$$

Assuming that $\hat{\gamma}_d$ converges to its true value, we can derive the relative nominal wage of each province compared to Zhejiang using the estimated fixed effects as follows:

$$\frac{I_d}{I_{\text{Zhejiang}}} = \mathrm{e}^{\hat{\gamma}_d}. \tag{E5}$$

We report these relative wage levels in Column (2) of Table E.1. Further, the relative nominal wages for any two provinces can be constructed as:

$$\frac{I_d}{I_o} = \frac{I_d}{I_{\text{Zhejiang}}} \frac{I_{\text{Zhejiang}}}{I_o} = \mathrm{e}^{\hat{\gamma}_d - \hat{\gamma}_o}. \tag{E6}$$

Table E.1 Regional fixed effects estimated by Liu (2024) and relative wage levels of various provinces.

	Region fixed effects in Liu (2024)	Relative wage level
Region	(1)	(2)
Zhejiang (reference)	0	1
Shandong	-0.168	0.845
Henan	-0.272	0.761
Jiangsu	-0.081	0.923
Anhui	-0.220	0.802
Jiangxi	-0.212	0.809
Guangdong	-0.150	0.861
Fujian	-0.171	0.843
Beijing	0.134	1.144
North Zhili	-0.215	0.806
Huguang (Hunan & Hubei)	-0.399	0.671

Notes: Column (1) replicates the estimated region fixed effects from Liu (2024) , with Zhejiang serving as the reference group. Column (2) reports the relative wage level of each province compared to Zhejiang, calculated based on the fixed effects. Italicized text denotes regions that are not included in our estimation, as they cannot be specifically matched to individual provinces in our data.

If we ignore the differences in price levels across provinces, we can use the above relative nominal wages to substitute for the relative real income $(\bar{I}_d/P_d)/(\bar{I}_o/P_o)$ required for estimating Equation (E3). Additionally, we can use these relative nominal wages to replace $\bar{I}_d/P_d$ in Equation (E1), given that the inclusion of destination fixed effects has accounted for the denominator.

Since the estimates provided by Liu (2024) are limited to eight provinces in our dataset, we restrict our analysis to the 64 province pairs that involve these provinces. We only use the migrant samples from the first sample period (i.e., the 1760s), as it is the closest period to the sample period in Liu (2024).

The estimation results are summarized in Table E.2. Panel A presents the estimates for Equation (E1), while Panel B reports the results for Equation (E3). Column (1) employs the PPML estimator, producing two estimates close to 3. In Column (2), the OLS estimator with log-linearization is used, reducing the sample size to 48. The resulting estimates diverge, with values of 3.89 and 2.19, respectively.

We acknowledge the potential endogeneity concern; specifically, the orthogonal condition that the error term is uncorrelated with the income term may not hold. However, obtaining historical income data is already challenging, let alone finding an appropriate instrumental variable for income. As a second-best approach, we construct an internal instrumental variable using the available income data. Following the methodology of Tombe and Zhu (2019), we use the distance-weighted average income of all other provinces as an instrumental variable for the income level of each destination province. The underlying rationale is that a province surrounded by neighbors with high income is likely to exhibit higher income itself, while the income levels of neighboring regions are plausibly exogenous to the migration and income shocks experienced by that province. The estimated results are presented in Column (3) of Table E.2, where we observe a smaller coefficient of approximately 1.64. Although the limited sample size and the inflated standard errors

commonly associated with instrumental variable estimation render these results statistically insignificant (Lal et al. 2024), the estimate falls within the range documented in studies on contemporary migration (as discussed in the main text). Thus, we believe they still provides a plausible estimate for κ. Taken together, our point estimates suggest that the shape parameter, which governs the dispersion of individual preferences, ranges between 1.64 and 3.89 during the Qing Dynasty.

Table E.2 Estimating the shape parameter κ.

	PPML	OLS (log-linearization)	IV (Poisson model)
	(1)	(2)	(3)
Panel A. Estimating Equation (E1)			
Relative wage of destination province	2.998**	3.888*	1.640
	(1.296)	(1.928)	(2.730)
Crossing *xiaqu* boundary	-0.574	-0.476	-0.573
	(0.459)	(0.453)	(0.475)
Logarithm of distance	-0.817***	-0.807***	-0.811***
	(0.061)	(0.114)	(0.058)
Clan difference	-1.717	-3.559*	-1.685
	(1.602)	(1.819)	(1.519)
Constant	0.458**	0.594	0.327
	(0.206)	(0.466)	(0.594)
Origin fixed effects	Yes	Yes	Yes
# of province-pair observations	64	48	64
Panel B. Estimating Equation (E3)			
Relative wage of destination province	2.999**	2.187*	1.652
	(1.301)	(1.138)	(2.118)
Crossing *xiaqu* boundary	-0.576	-0.487	-0.575*
	(0.463)	(0.396)	(0.310)
Logarithm of distance	-0.819***	-0.822***	-0.813***
	(0.061)	(0.102)	(0.088)
Clan difference	-1.653	-3.092	-1.625
	(1.623)	(2.366)	(2.557)
Constant	-0.000	0.012	-0.000
	(0.000)	(0.058)	(0.001)
# of province-pair observations	64	48	64

Notes: This table presents the estimates for the shape parameter κ. Panel A corresponds to the estimation of Equation (E1), while Panel B corresponds to Equation (E3). Column (1) reports results obtained using the PPML estimator, while Column (2) uses the OLS estimator with log-linearization applied to the dependent variable. Column (3) provides the instrumental variable estimation using the Poisson model, where the instrument is the distance-weighted average income of all other provinces. Standard errors are reported in parentheses.
***p < 0.01; **p < 0.05; *p < 0.1.

F Supplementary figures and tables

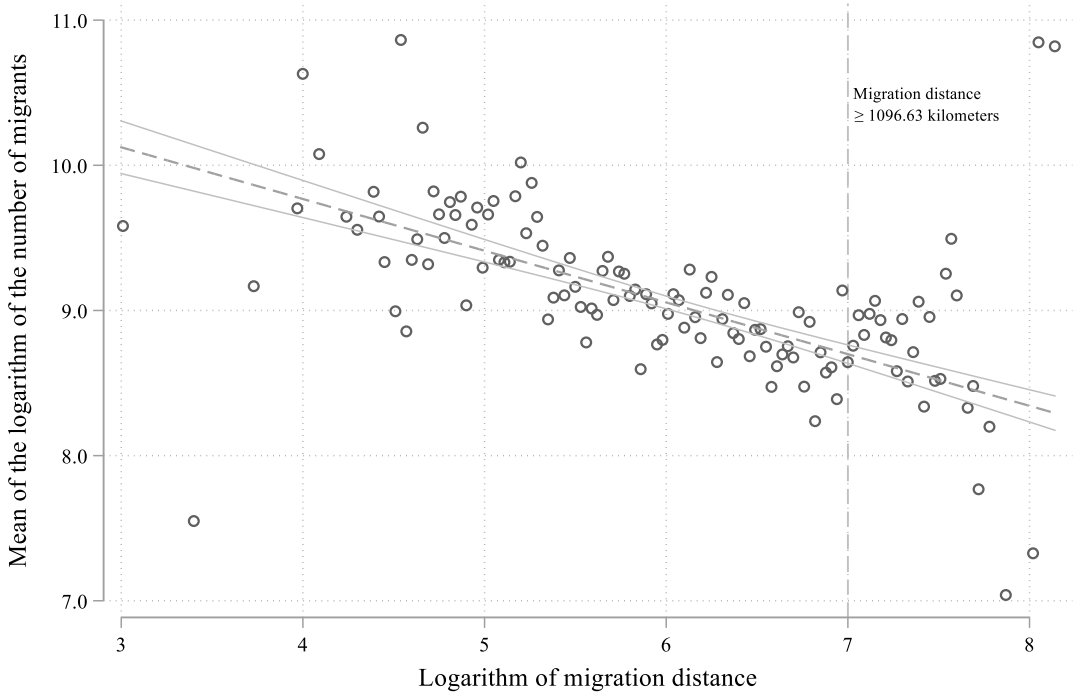

Figure F.1 The relationship between the number of migrants and migration distance.
Notes: This figure is a binscatter depicting the mean logarithm of the number of migrants in relation to the distance between two prefectures. Each dot represents the mean value for all prefecture pairs within a bin of width 0.03. As the number of migrants is logarithmically transformed to align to the gravity equation in the main text, prefecture pairs with zero migrants are excluded from the graph. In other words, this graph illustrates the relationship between migration and physical distance, conditional on the existence of migrants.

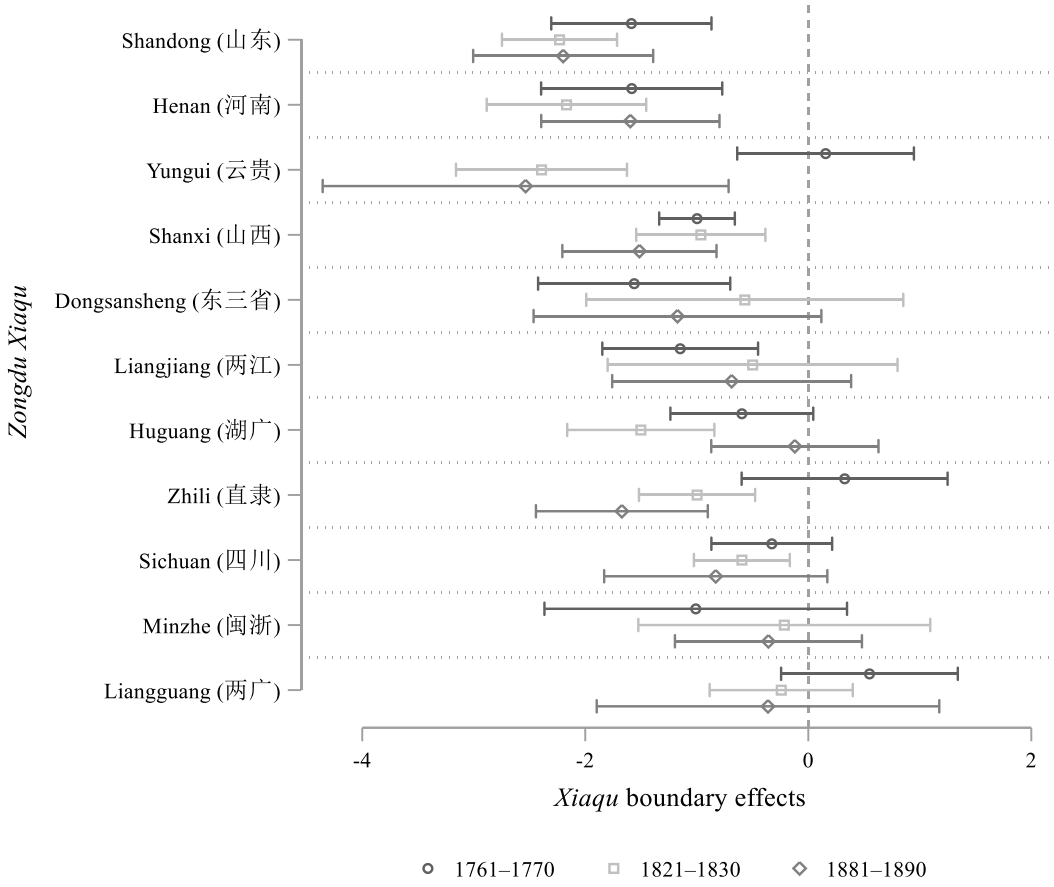

Figure F.2 Heterogeneous *xiaqu* boundary effects, by period.
Notes: This figure reports the estimated boundary effects for different destination *xiaqu* by period, derived from estimating Equation (13) with province-pair datasets for different periods. The Neimenggu (内蒙古) *xiaqu* is not included in this figure, as there were too few recorded in-migrants to provide accurate estimates. The Shangan (陕甘) *xiaqu* is also excluded, as we observe distinctly large and dubious boundary effects for its third period. The table version of these estimates is reported in Table F.13. Point estimates are represented by dots, while 90% confidence intervals are depicted as lines, calculated using robust standard errors two-way clustering at the origin and destination provinces.

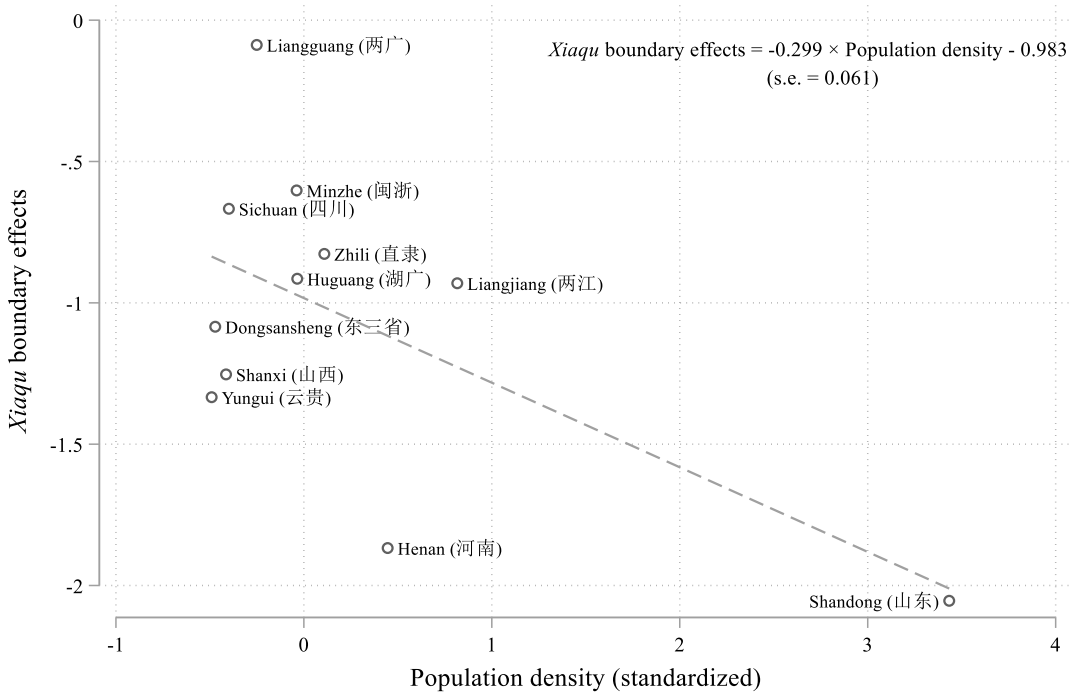

Figure F.3 Correlations between *xiaqu* boundary effects with population density.
Notes: This figure illustrates the correlation between the estimated *xiaqu* boundary effects and population density. The boundary effect for each *xiaqu* is derived from estimating Equation (13) and is presented in Table F.12. Population density is first averaged across three periods and then standardized. The results from a univariate regression are indicated in the figure, with standard errors robust to heteroskedasticity reported in parentheses.

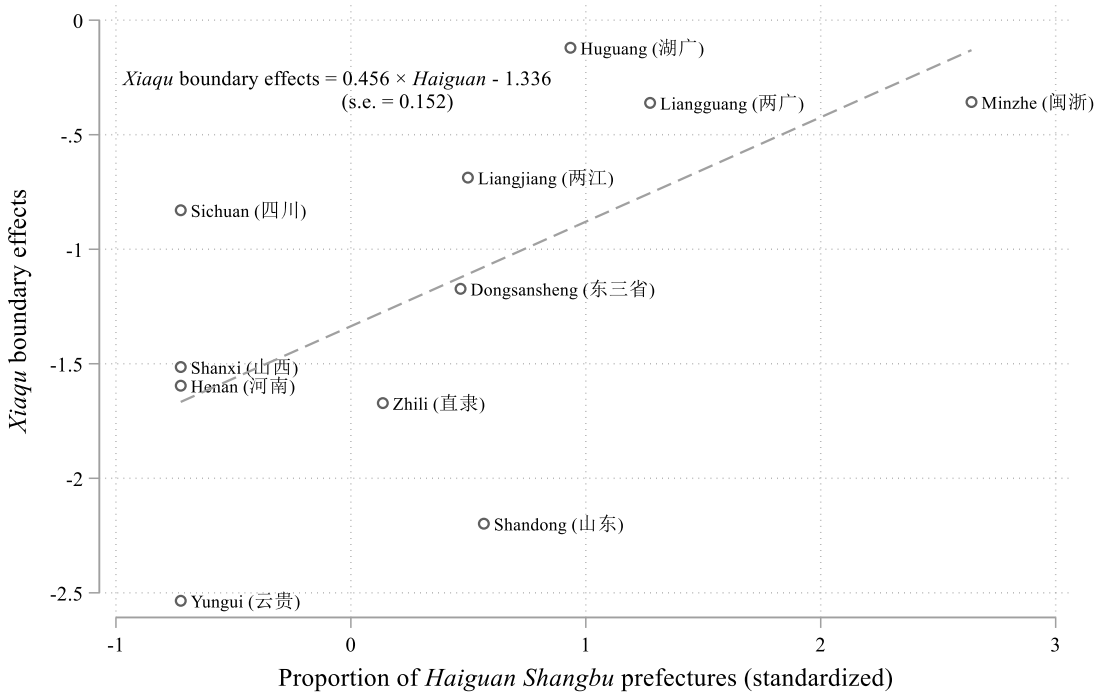

Figure F.4 Correlations between *xiaqu* boundary effects with the degree of trade openness.
Notes: This figure illustrates the correlation between the estimated *xiaqu* boundary effects and the degree of trade openness, with the latter measured by the proportion of *Haiguan Shangbu* (海关商埠) prefectures within each *xiaqu*. Since the *Haiguan Shangbu* was established around the third sample period (i.e., the 1880s), this figure presents their relationship only for that period. Specifically, the *xiaqu* boundary effects on the y-axis are those estimated using data from the third period, as shown in Table F.13. The proportion on the x-axis is standardized. The results from a univariate regression are indicated in the figure, with standard errors robust to heteroskedasticity reported in parentheses.

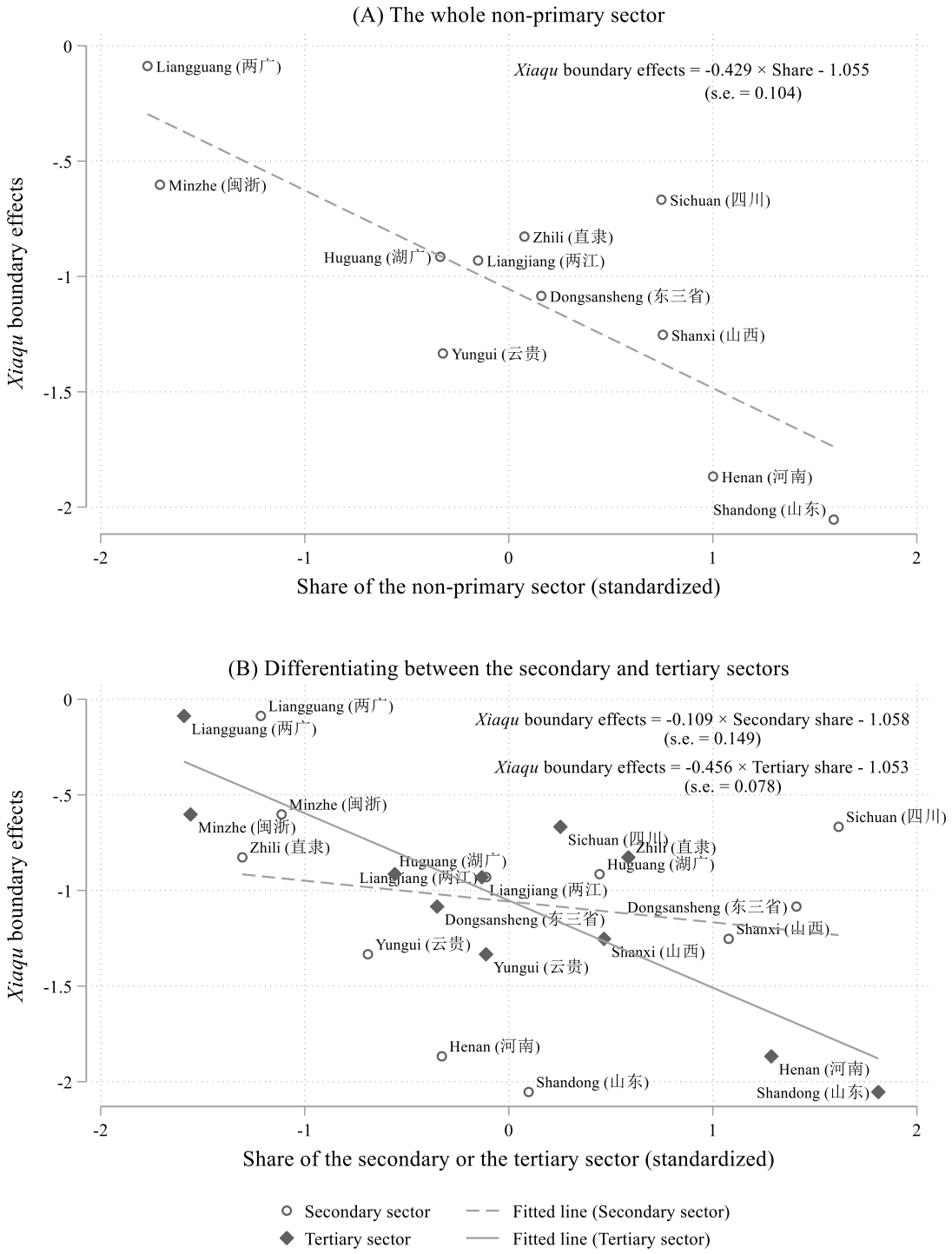

Figure F.5 Correlations between *xiaqu* boundary effects with the occupation structure.
Notes: This figure illustrates the correlation between the estimated *xiaqu* boundary effects and the occupation structure. The boundary effect for each *xiaqu* is derived from estimating Equation (13) and is presented in Table F.12. Panel (A) examines the share of the entire non-primary sector (i.e., the secondary and tertiary sectors), while Panel (B) separately presents the share of the secondary and tertiary sectors. All shares are first averaged across three periods and then standardized. The results from several univariate regressions are indicated in the figure, with standard errors robust to heteroskedasticity reported in parentheses.

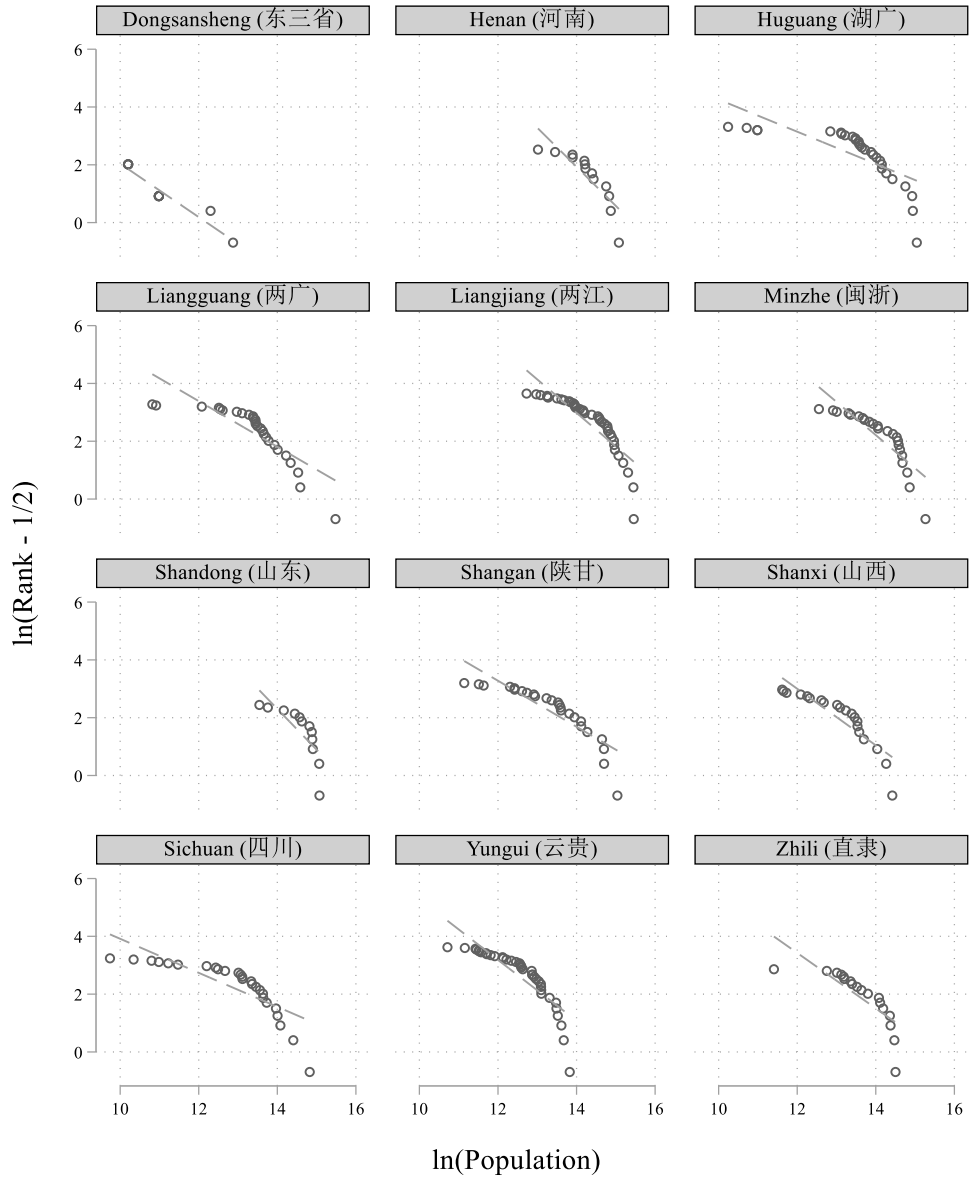

Figure F.6 *Xiaqu*-level fits of the Zipf's law, 1761–1770.
Notes: This figure shows the fits of the Zipf's law at the *xiaqu* level with population data in the first sample period (i.e., 1760s).

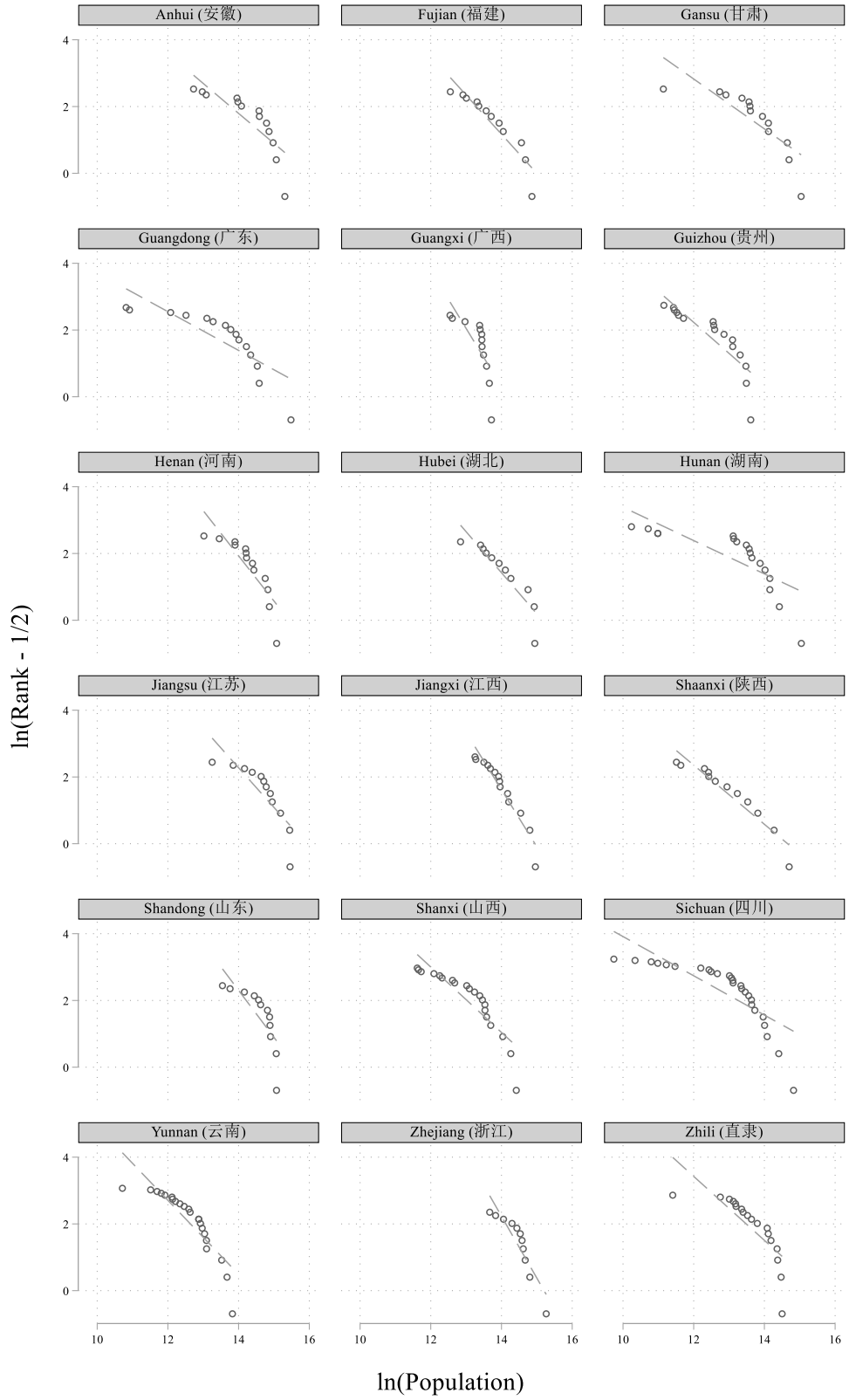

Figure F.7 Province-level fits of the Zipf's law, 1761–1770.
Notes: This figure shows the fits of the Zipf's law at the province level with population data in the first sample period (i.e., 1760s).

Table F.1 Administrative divisions involved in this study.

Zongdu Xiaqu	Province	Prefecture
Dongsansheng (东三省)	Heilongjiang (黑龙江)	Hulunbeier Fudutong Xiaqu (呼伦贝尔副都统辖区), Heilongjiang Fudutong Xiaqu (黑龙江副都统辖区), Moergen Fudutong Xiaqu (墨尔根副都统辖区), Qiqihaer Fudutong Xiaqu (齐齐哈尔副都统辖区)
	Jilin (吉林)	Sanxing Fudutong Xiaqu (三姓副都统辖区), Ningguta Fudutong Xiaqu (宁古塔副都统辖区), Alechuka Fudutong Xiaqu (阿勒楚喀副都统辖区), Baidune Fudutong Xiaqu (白都纳副都统辖区), Jilin Fudutong Xiaqu (吉林副都统辖区)
	Shengjing (盛京)	Fengtian Fu (奉天府), Yangxi Muchang (养息牧场), Dalinghe Muchang (大凌河牧场), Jinzhou Fu (锦州府)
Henan (河南)	Henan (河南)	Xu Zhou (许州), Nanyang Fu (南阳府), Chenzhou Fu (陈州府), Ru Zhou (汝州), Henan Fu (河南府), Huaiqing Fu (怀庆府), Zhangde Fu (彰德府), Weihui Fu (卫辉府), Kaifeng Fu (开封府), Guide Fu (归德府), Guang Zhou (光州), Runing Fu (汝宁府), Shan Zhou (陕州)
Huguang (湖广)	Hubei (湖北)	Yunyang Fu (郧阳府), Xiangyang Fu (襄阳府), Dean Fu (德安府), Hanyang Fu (汉阳府), Huangzhou Fu (黄州府), Wuchang Fu (武昌府), Anlu Fu (安陆府), Jingmen Zhou (荆门州), Yichang Fu (宜昌府), Jingzhou Fu (荆州府), Shinan Fu (施南府)
	Hunan (湖南)	Yongsui Ting (永绥厅), Qianzhou Ting (乾州厅), Fenghuang Ting (凤凰厅), Huangzhou Ting (晃州厅), Li Zhou (澧州), Changde Fu (常德府), Yuezhou Fu (岳州府), Changsha Fu (长沙府), Baoqing Fu (宝庆府), Yongshun Xian (永顺府), Chenzhou Fu (辰州府), Yuanzhou Fu (沅州府), Jing Zhou (靖州), Hengzhou Fu (衡州府), Yongzhou Fu (永州府), Guiyang Zhou (桂阳州), Chen Zhou (郴州)
Liangguang (两广)	Guangdong (广东)	Qiongzhou Fu (琼州府), Wanlichangsha (万里长沙), Qianlishitang (千里石塘), Dongsha (东沙), Fogang Ting (佛冈厅), Lianshan Ting (连山厅), Shaozhou Fu (韶州府), Nanxiong Zhou (南雄州), Lian Zhou (连州), Huizhou Fu (惠州府), Jiaying Zhou (嘉应州), Guangzhou Fu (广州府), Gaozhou Fu (高州府), Luoding Zhou (罗定州), Leizhou Fu (雷州府), Lianzhou Fu (廉州府), Zhaoqing Fu (肇庆府), Chaozhou Fu (潮州府)
	Guangxi (广西)	Liuzhou Fu (柳州府), Pingle Fu (平乐府), Zhenan Fu (镇安府), Sien Fu (思恩府), Qingyuan Fu (庆远府), Sicheng Fu (泗城府), Guilin Fu (桂林府), Nanning Fu (南宁府), Taiping Fu (太平府), Yulin Zhou (郁林州), Xunzhou Fu (浔州府), Wuzhou Fu (梧州府)
Liangjiang (两江)	Anhui (安徽)	Huizhou Fu (徽州府), Anqing Fu (安庆府), Chizhou Fu (池州府), Ningguo Fu (宁国府), Luzhou Fu (庐州府), Taiping Fu (太平府（皖）), Fengyang Fu (凤阳府), Yingzhou Fu (颍州府), Si Zhou (泗州), Chu Zhou (滁州), He Zhou (和州), Guangde Zhou (广德州), Liuan Zhou (六安州)
	Jiangsu (江苏)	Zhenjiang Fu (镇江府), Changzhou Fu (常州府), Taicang Zhou (太仓州), Songjiang Fu (松江府), Xuzhou Fu (徐州府), Hai Zhou (海州), Huaian Fu (淮安府), Yangzhou Fu (扬州府), Jiangning Fu (江宁府), Haimen Ting (海门厅), Suzhou Fu (苏州府), Tong Zhou (通州)
	Jiangxi (江西)	Zhenjiang Fu (镇江府), Changzhou Fu (常州府), Taicang Zhou (太仓州), Songjiang Fu (松江府), Xuzhou Fu (徐州府), Hai Zhou (海州), Huaian Fu (淮安府), Yangzhou Fu (扬州府), Jiangning Fu (江宁府), Haimen Ting (海门厅), Suzhou Fu (苏州府), Tong Zhou (通州)
Minzhe (闽浙)	Fujian (福建)	Taiwan Fu (台湾府), Zhangzhou Fu (漳州府), Tingzhou Fu (汀州府), Quanzhou Fu (泉州府), Longyan Zhou (龙岩州), Xinghua Fu (兴化府), Fuzhou Fu (福州府), Funing Fu (福宁府), Jianning Fu (建宁府), Shaowu Fu (邵武府), Yanping Fu (延平府), Yongchun Zhou (永春州)
	Zhejiang (浙江)	Jiaxing Fu (嘉兴府), Huzhou Fu (湖州府), Wenzhou Fu (温州府), Chuzhou Fu (处州府), Hangzhou Fu (杭州府), Yanzhou Fu (严州府), Taizhou Fu (台州府), Quzhou Fu (衢州府), Jinhua Fu (金华府), Shaoxing Fu (绍兴府), Ningpo Fu (宁波府)

Table F.1 (continued) Administrative divisions involved in this study.

Zongdu Xiaqu	Province	Prefecture
Neimenggu (内蒙古)	Neimenggu (内蒙古)	Zhelimu Meng (哲里木盟), Daligangai Muchang (达里冈爱牧场), Ejinatuerhute Qi (额济纳土尔扈特旗), Xilinguole Meng (锡林郭勒盟), Alashanelute Qi (阿拉善厄鲁特旗), Wulanchabu Meng (乌兰察布盟), Zhaowuda Meng (昭乌达盟), Yikezhao Meng (伊克昭盟)
Shandong (山东)	Shandong (山东)	Qingzhou Fu (青州府), Taian Fu (泰安府), Laizhou Fu (莱州府), Dengzhou Fu (登州府), Linqing Zhou (临清州), Dongchang Fu (东昌府), Wuding Fu (武定府), Jinan Fu (济南府), Yizhou Fu (沂州府), Yanzhou Fu (兖州府), Jining Zhou (济宁州), Caozhou Fu (曹州府)
Shangan (陕甘)	Gansu (甘肃)	Liangzhou Fu (凉州府), Ganzhou Fu (甘州府), Anxi Zhou (安西州), Su Zhou (肃州), Ningxia Fu (宁夏府), Xining Fu (西宁府), Lanzhou Fu (兰州府), Gongchang Fu (巩昌府), Pingliang Zhou (平凉府), Jin Zhou (泾州), Qingynag Fu (庆阳府), Jie Zhou (阶州), Qin Zhou (秦州)
	Shaanxi (陕西)	Yulin Fu (榆林府), Suide Zhou (绥德州), Yanan Fu (延安府), Fu Zhou (鄜州), Tong Zhou (同州府), Shang Zhou (商州), Xingan Fu (兴安府), Xian Fu (西安府), Fengxiang Fu (凤翔府), Bin Zhou (邠州), Qian Zhou (乾州), Hanzhong Fu (汉中府)
	Xinjiang (新疆)	Yeerqiang (叶尔羌), Akesu (阿克苏), Wulumuqi (乌噜木齐), Gucheng (古城), Tulufan (吐鲁番), Balikun (巴里坤), Hami (哈密), Hetian (和阗), Kuche (库车), Kalashaer (喀喇沙尔), Wushen (乌什), Kuerkalawu (喀喇乌苏), Geshengeer (喀什噶尔), Taerbahatai (塔尔巴哈台), Yili (伊犁)
Shanxi (山西)	Shanxi (山西)	Xin Zhou (忻州), Qin Zhou (沁州), Huo Zhou (霍州), Liao Zhou (辽州), Luan Fu (潞安府), Zezhou Fu (泽州府), Pingyang Fu (平阳府), Jie Zhou (解州), Xi Zhou (隰州), Fenzhou Fu (汾州府), Ningwu Fu (宁武府), Taiyuan Fu (太原府), Baode Zhou (保德州), GuiSuiLiTing (归绥六厅), Shuoping Fu (朔平府), Datong Fu (大同府), Dai Zhou (代州), Pingding Zhou (平定州), Jiang Zhou (绛州), Puzhou Fu (蒲州府)
Sichuan (四川)	Sichuan (四川)	Zhong Zhou (忠州), Xuzhou Fu (叙州府), Jiading Fu (嘉定府), Songpan Ting (松潘厅), Zagu Ting (杂谷厅), Maogong Ting (懋功厅), Yazhou Fu (雅州府), Ningyuan Fu (宁远府), Qiong Zhou (邛州), Mei Zhou (眉州), Mao Zhou (茂州), Lu Zhou (泸州), Chongqing Fu (重庆府), Longan Fu (龙安府), Chengdu Fu (成都府), Mian Zhou (绵州), Zi Zhou (资州), Tongzhou Fu (潼川府), Shunqing Fu (顺庆府), Suiding Fu (绥定府), Xuyong Ting (叙永厅), Shizhu Ting (石砫厅), Baoning Fu (保宁府), Youyang Zhou (酉阳州), Kuizhou Fu (夔州府), Taiping Ting (太平厅)
Yungui (云贵)	Guizhou (贵州)	Renhuai Ting (仁怀厅), Zunyi Fu (遵义府), Dading Fu (大定府), Puan Ting (普安厅), Anshun Fu (安顺府), Xingyi Fu (兴义府), Guiyang Fu (贵阳府), Pingyue Zhou (平越州), Duyun Fu (都匀府), Sinan Fu (思南府), Shiqian Fu (石阡府), Songtao Ting (松桃厅), Tongren Fu (铜仁府), Sizhou Fu (思州府), Zhenyuan Fu (镇远府), Liping Fu (黎平府)
	Yunnan (云南)	Kaihua Fu (开化府), Zhaotong Fu (昭通府), Dongchuan Fu (东川府), Yongbei Ting (永北厅), Wuding Zhou (武定州), Qujing Fu (曲靖府), Chuxiong Fu (楚雄府), Yunnan Fu (云南府), Yuanjiang Zhou (元江州), Chengjiang Fu (澂江府), Linan Fu (临安府), Lijiang Fu (丽江府), Tengyue Ting (腾越厅), Dali Fu (大理府), Yongchang Fu (永昌府), Menghua Ting (蒙化厅), Shunning Fu (顺宁府), Jingdong Ting (景东厅), Zhenyuan Zhou (镇沅州), Puer Fu (普洱府), Guangxi Zhou (广西州), Guangnan Fu (广南府)
Zhili (直隶)	Zhili (直隶)	Daming Fu (大名府), Guangping Fu (广平府), Shunde Fu (顺德府), Zhao Zhou (赵州), Ji Zhou (冀州), Shen Zhou (深州), Hejian Fu (河间府), Baoding Fu (保定府), Ding Zhou (定州), Zhengding Fu (正定府), Tianjin Fu (天津府), Shuntian Fu (顺天府), Yi Zhou (易州), Xuanhua Fu (宣化府), Zunhua Zhou (遵化州), Yongping Fu (永平府), Chengde Fu (承德府), Koubeisanting (口北三厅)

Table F.2 Comparison between the PPML estimator and OLS estimator.

	PPML	OLS with log-linearization
	(1)	(2)
Crossing *xiaqu* boundary	-0.372***	-0.047
	(0.123)	(0.066)
Crossing province boundary	0.279**	0.030
	(0.116)	(0.068)
Origin fixed effects × Period fixed effects	Yes	Yes
Destination fixed effects × Period fixed effects	Yes	Yes
Controls × Period fixed effects	Yes	Yes
# of migrants in *XKTB*	9,259	8,882
# of prefecture-pair observations	166,329	3,564

Notes: This table presents estimation results using the PPML estimator and OLS estimator. Column (1) replicates the specification of Column (5) from Table 2, while Columns (2) estimates the following equation using OLS:

$$\ln m_{odt,-o} = \ln\gamma + \alpha_o \times \delta_t + \beta_d \times \delta_t - \sum_t \sigma_t \cdot \ln Distance_{od} \times \delta_t + \sum_t \pi_t (1 - Culture_{odt}) \times \delta_t + \left(\ln \bar{b}^p\right) \mathbb{I}\{Prov_o \neq Prov_d\} + \left(\ln \bar{b}^x\right) \mathbb{I}\{Xiaqu_o \neq Xiaqu_d\} + \varepsilon_{odt},$$

where α_o, β_d, and δ_t denote origin, destination, and period fixed effects, respectively. In other words, all fixed effects and variables, except the two dummies indicating administrative boundary crossings, interact with period fixed effects. Thus, Columns (2) follows similar settings to Column (1). The dependent variable $m_{odt,-o}$ represents either the migration share itself or a variant of its log-like transformation. All regressions include constant terms that are not listed in the table. Robust standard errors, two-way clustered at the origin and destination prefectures, are reported in parentheses.
***p < 0.01; **p < 0.05; *p < 0.1.

Table F.3 Standard errors clustering at different levels.

	Dependent variable: Migrant share among all out-migrants				
	1760s	1820s	1880s	All three periods	
	(1)	(2)	(3)	(4)	(5)
Crossing *xiaqu* boundary	-0.396**	-0.398**	-0.312	-0.326**	-0.372**
	(0.173)	(0.168)	(0.246)	(0.114)	(0.123)
	{0.191}	{0.162}	{0.244}	{0.141}	{0.152}
Crossing province boundary	0.277	0.298*	0.258	0.257**	0.279**
	(0.175)	(0.169)	(0.237)	(0.104)	(0.116)
	{0.182}	{0.135}	{0.275}	{0.126}	{0.137}
Logarithm of distance	-2.205***	-2.520***	-2.404***	-2.340***	
	(0.101)	(0.095)	(0.112)	(0.059)	
	{0.100}	{0.097}	{0.145}	{0.069}	
Clan difference	-1.468**	-1.083*	-0.424	-0.434***	
	(0.572)	(0.548)	(0.558)	(0.166)	
	{0.556}	{0.608}	{0.649}	{0.155}	
Constant	9.269***	10.880***	10.061***	9.405***	10.077***
	(0.527)	(0.494)	(0.636)	(0.319)	(0.354)
	{0.571}	{0.556}	{0.835}	{0.399}	{0.483}
Origin fixed effects	Yes	Yes	Yes	Yes	No
Destination fixed effects	Yes	Yes	Yes	Yes	No
Period fixed effects	No	No	No	Yes	No
Origin FEs × Period FEs	No	No	No	No	Yes
Destination FEs × Period FEs	No	No	No	No	Yes
Controls × Period FEs	No	No	No	No	Yes
# of migrants in *XKTB*	4,027	3,393	1,839	9,259	9,259
# of prefecture-pair observations	58,566	62,224	45,539	186,682	166,329
# of clusters: origin prefectures	238	243	205	267	267
# of clusters: destination prefectures	247	257	233	277	277
# of clusters: origin provinces	20	20	21	21	21
# of clusters: destination provinces	22	21	22	23	23

Notes: This table provides estimation results using the PPML estimator. Robust standard errors two-way clustering at the origin and destination prefectures, are shown in parentheses; robust standard errors two-way clustering at the origin and destination provinces, are shown in braces.
***p < 0.01; **p < 0.05; *p < 0.1. Marked based on the largest p-value.

Table F.4 Estimations without long-distance migration.

	Dependent variable: Migrant share among all non-long-distance out-migrants		
	All pairs	Excl. inter-province, intra-*xiaqu* outliers	Excl. inter-province outliers
	(1)	(2)	(3)
Crossing *xiaqu* boundary	-0.330**	-0.226**	-0.321***
	(0.152)	(0.115)	(0.113)
Crossing province boundary	0.288**	0.153	0.087
	(0.143)	(0.116)	(0.111)
Origin FEs × Period FEs	Yes	Yes	Yes
Destination FEs × Period FEs	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes
# of migrants in *XKTB*	8,893	8,873	8,819
# of prefecture-pair observations	95,422	93,859	89,314

Notes: This table examines the robustness of the results concerning long-distance migrants, partly influenced by specific policies aimed at promoting migration and reclamation. All estimations use the PPML estimator. The dependent variable is the migrant share among all out-migrants originating from the origin prefecture in each prefecture pair, restricted to migration distances shorter than 1,100 kilometers. Column (2) excludes inter-province, intra-*xiaqu* outlier pairs, while Column (3) excludes all inter-province outlier pairs, as detailed in Table F.8. All regressions include constant terms, which are not reported in the table. Robust standard errors, clustered two-way at the origin and destination prefectures, are provided in parentheses. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.5 Estimations using migrant shares without reweighting as the dependent variable.

	Dependent var.: Migrant share among all out-migrants without reweighting		
	All pairs	Excl. inter-province, intra-*xiaqu* outliers	Excl. inter-province outliers
	(1)	(2)	(3)
Crossing *xiaqu* boundary	-0.335***	-0.167‡	-0.264**
	(0.128)	(0.118)	(0.111)
Crossing province boundary	0.241*	0.055	0.007
	(0.128)	(0.124)	(0.117)
Origin FEs × Period FEs	Yes	Yes	Yes
Destination FEs × Period FEs	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes
# of migrants in *XKTB*	9,260	9,238	9,178
# of prefecture-pair observations	166,534	162,905	153,306

Notes: This table presents estimation results using the PPML estimator. The dependent variable is constructed from the raw number of migrants recorded in *XTKB*, without reweighting for the ratio of historically documented resident numbers to *XTKB* individual numbers. Column (2) excludes inter-province, intra-*xiaqu* outlier pairs, while Column (3) excludes all inter-province outlier pairs, as detailed in Table F.8. All regressions include constant terms, which are not reported in the table. Robust standard errors, with two-way clustering at the origin and destination prefectures, are shown in parentheses.
***p < 0.01; **p < 0.05; *p < 0.1; ‡p < 0.16.

Table F.6 Estimations with the number of migrants as the dependent variable.

	All pairs	Excl. inter-province, intra-*xiaqu* outliers	Excl. inter-province outliers
	(1)	(2)	(3)
Panel A. Dependent variable: number of migrants in XKTB			
Crossing *xiaqu* boundary	-0.412***	-0.393**	-0.408***
	(0.118)	(0.180)	(0.118)
Crossing province boundary	0.316**	0.295	0.285**
	(0.136)	(0.206)	(0.138)
# of migrants in *XKTB*	9,260	9,238	9,178
# of prefecture-pair observations	166,534	162,905	153,306
Panel B. Dependent variable: reweighted number of migrants			
Crossing *xiaqu* boundary	-0.394***	-0.371*	-0.389***
	(0.133)	(0.204)	(0.135)
Crossing province boundary	0.290**	0.261	0.249*
	(0.138)	(0.199)	(0.140)
# of migrants in *XKTB*	9,259	9,237	9,177
# of prefecture-pair observations	166,329	162,705	153,121
Origin FEs × Period FEs	Yes	Yes	Yes
Destination FEs × Period FEs	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes

Notes: This table presents estimation results using the PPML estimator. The dependent variable in Panel A is the raw number of migrants recorded in *XKTB*, and that in Panel B is the reweighted number of migrants, calculated based on the ratio of historically documented resident numbers to *XTKB* individual numbers. Column (2) excludes inter-province, intra-*xiaqu* outlier pairs, while Column (3) excludes all inter-province outlier pairs, as detailed in Table F.8. All regressions include constant terms, which are not reported in the table. Robust standard errors, with two-way clustering at the origin and destination prefectures, are shown in parentheses.

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.7 Estimations with different specifications of migration distances.

	Dependent variable: Migrant share among all out-migrants					
	(1)	(2)	(3)	(4)	(5)	(6)
Crossing *xiaqu* boundary	-0.372***	-0.284**	-0.338***	-0.366***	-0.390***	-0.391***
	(0.123)	(0.118)	(0.123)	(0.124)	(0.123)	(0.123)
Crossing prov. boundary	0.279**	-0.0648	0.228*	0.289**	0.326***	0.328***
	(0.116)	(0.119)	(0.118)	(0.117)	(0.117)	(0.117)
Logarithm of distance × $Period_{1760s}$	-2.213***					
	(0.092)					
Logarithm of distance × $Period_{1820s}$	-2.522***					
	(0.085)					
Logarithm of distance × $Period_{1880s}$	-2.391***					
	(0.100)					
Distance × $Period_{1760s}$		-0.008***	-0.013***	-0.016***	-0.021***	-0.024***
		(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Distance × $Period_{1820s}$		-0.009***	-0.017***	-0.020***	-0.031***	-0.030***
		(0.001)	(0.002)	(0.002)	(0.003)	(0.005)
Distance × $Period_{1880s}$		-0.008***	-0.015***	-0.018***	-0.021***	-0.022***
		(0.001)	(0.001)	(0.001)	(0.002)	(0.003)
$Distance^2$ × $Period_{1760s}$		2.29e-06***	7.90e-06***	1.31e-05***	2.51e-05***	3.61e-05***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$Distance^2$ × $Period_{1820s}$		2.31e-06***	1.35e-05***	2.09e-05***	5.36e-05***	4.79e-05**
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$Distance^2$ × $Period_{1880s}$		1.73e-06***	1.03e-05***	1.70e-05***	2.48e-05***	2.74e-05***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$Distance^3$ × $Period_{1760s}$			-1.47e-09***	-4.39e-09***	-1.52e-08***	-2.95e-08***
			(0.000)	(0.000)	(0.000)	(0.000)
$Distance^3$ × $Period_{1820s}$			-3.74e-09***	-9.44e-09***	-4.81e-08***	-3.73e-08
			(0.000)	(0.000)	(0.000)	(0.000)
$Distance^3$ × $Period_{1880s}$			-2.38e-09***	-6.81e-09***	-1.43e-08***	-1.79e-08*
			(0.000)	(0.000)	(0.000)	(0.000)
$Distance^4$ × $Period_{1760s}$				4.97e-13***	4.36e-12***	1.27e-11***
				(0.000)	(0.000)	(0.000)
$Distance^4$ × $Period_{1820s}$				1.37e-12***	2.00e-11***	1.04e-11
				(0.000)	(0.000)	(0.000)
$Distance^4$ × $Period_{1880s}$				9.05e-13***	3.74e-12***	6.04e-12
				(0.000)	(0.000)	(0.000)
$Distance^5$ × $Period_{1760s}$					-4.65e-16***	-2.70e-15**
					(0.000)	(0.000)
$Distance^5$ × $Period_{1820s}$					-3.05e-15***	8.26e-16
					(0.000)	(0.000)
$Distance^5$ × $Period_{1880s}$					-3.61e-16**	-1.02e-15
					(0.000)	(0.000)
$Distance^6$ × $Period_{1760s}$						2.18e-19*
						(0.000)
$Distance^6$ × $Period_{1820s}$						-5.85e-19
						(0.000)
$Distance^6$ × $Period_{1880s}$						6.94e-20
						(0.000)
Origin FEs × Period FEs	Yes	Yes	Yes	Yes	Yes	Yes
Dest. FEs × Period FEs	Yes	Yes	Yes	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes	Yes	Yes	Yes
# of migrants in *XKTB*	9,259	9,259	9,259	9,259	9,259	9,259
# of pref.-pair observations	166,329	166,329	166,329	166,329	166,329	166,329

Notes: This table tests the robustness regarding the specification of migration distance, a key confounding factor for estimating administrative boundary effects. All estimations use the PPML estimator and include constant terms that are not listed in the table. Robust standard errors, with two-way clustering at the origin and destination prefectures, are shown in parentheses.
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.8 List of outlier prefecture pairs.

Origin prefecture	Destination prefecture	Period	# of migrants in *XKTB*
Panel A. Inter-province, intra-xiaqu outlier pairs			
Yulin Zhou (郁林州)	Qiongzhou Fu (琼州府)	1761–1770	4
Xingyi Fu (兴义府)	Dongchuan Fu (东川府)	1761–1770	1
Puan Ting (普安厅)	Qujing Fu (曲靖府)	1761–1770	1
Jilin Fudutong Xiaqu (吉林副都统辖区)	Fengtian Fu (奉天府)	1761–1770	1
Chu Zhou (滁州)	Zhenjiang Fu (镇江府)	1761–1770	1
Zhenjiang Fu (镇江府)	Luzhou Fu (庐州府)	1821–1830	1
Lianzhou Fu (廉州府)	Nanning Fu (南宁府)	1821–1830	1
Fu Zhou (鄜州)	Qingyang Fu (庆阳府)	1821–1830	1
Jilin Fudutong Xiaqu (吉林副都统辖区)	Fengtian Fu (奉天府)	1821–1830	3
Jiangning Fu (江宁府)	Luzhou Fu (庐州府)	1881–1830	1
Su Zhou (肃州)	Tongzhou Fu (同州府)	1881–1830	1
Chu Zhou (滁州)	Jiangning Fu (江宁府)	1881–1830	2
Balikun (巴里坤)	Hanzhong Fu (汉中府)	1881–1830	1
Yanan Fu (延安府)	Ningxia Fu (宁夏府)	1881–1830	1
Bin Zhou (邠州)	Qingyang Fu (庆阳府)	1881–1830	2
Gucheng (古城)	Tongzhou Fu (同州府)	1881–1830	1
Panel B. Inter-xiaqu outlier pairs			
Haimen Ting (海门厅)	Shaoxing Fu (绍兴府)	1761–1770	1
Xingan Fu (兴安府)	Anshun Fu (安顺府)	1761–1770	1
Guangnan Fu (广南府)	Sicheng Fu (泗城府)	1761–1770	1
Yunyang Fu (郧阳府)	Shang Zhou (商州)	1761–1770	2
Kaihua Fu (开化府)	Shuntian Fu (顺天府)	1761–1770	1
Songjiang Fu (松江府)	Caozhou Fu (曹州府)	1761–1770	1
Guangxin Fu (广信府)	Jianning Fu (建宁府)	1761–1770	4
Lian Zhou (连州)	Shunqing Fu (顺庆府)	1761–1770	1
Ningyuan Fu (宁远府)	Qingyuan Fu (庆远府)	1761–1770	1
Puer Fu (普洱府)	Taicang Zhou (太仓州)	1761–1770	1
Zhaotong Fu (昭通府)	Xuzhou Fu (叙州府)	1761–1770	1
Sien Fu (思恩府)	Youyang Zhou (酉阳州)	1761–1770	1
Yanan Fu (延安府)	Xi Zhou (隰州)	1761–1770	3
Sicheng Fu (泗城府)	Guangnan Fu (广南府)	1761–1770	2
Taiping Fu (太平府)	Jianning Fu (建宁府)	1761–1770	1
Qiongzhou Fu (琼州府)	Mei Zhou (眉州)	1821–1830	1
Xi Zhou (隰州)	Baoning Fu (保宁府)	1821–1830	1
Liping Fu (黎平府)	Fengtian Fu (奉天府)	1821–1830	1
Taicang Zhou (太仓州)	Jiaxing Fu (嘉兴府)	1821–1830	3
Pingle Fu (平乐府)	Yongzhou Fu (永州府)	1821–1830	2
Yanan Fu (延安府)	Xuanhua Fu (宣化府)	1821–1830	1
Taiping Fu (太平府)	Yichang Fu (宜昌府)	1821–1830	1
Guangxin Fu (广信府)	Quzhou Fu (衢州府)	1821–1830	2
Yanping Fu (延平府)	Huizhou Fu (惠州府)	1821–1830	1
Renhuai Ting (仁怀厅)	Xuyong Ting (叙永厅)	1821–1830	1
Ningdu Zhou (宁都州)	Jianning Fu (建宁府)	1821–1830	4
Koubeisanting (口北三厅)	Datong Fu (大同府)	1881–1830	1
Jiang Zhou (绛州)	Nanyang Fu (南阳府)	1881–1830	1
Lian Zhou (连州)	Xuzhou Fu (徐州府)	1881–1830	1
Yulin Fu (榆林府)	Guisuiliuting (归绥六厅)	1881–1830	5
Huangzhou Ting (晃州厅)	Zhenyuan Fu (镇远府)	1881–1830	1
Yichang Fu (宜昌府)	Shuoping Fu (朔平府)	1881–1830	1
Yongbei Ting (永北厅)	Zi Zhou (资州)	1881–1830	1
Guangde Zhou (广德州)	Huzhou Fu (湖州府)	1881–1830	3
Liuan Zhou (六安州)	Xian Fu (西安府)	1881–1830	1

Table F.8 (continued) List of outlier prefecture pairs.

Origin prefecture	Destination prefecture	Period	# of migrants in *XKTB*
Fu Zhou (鄜州)	Datong Fu (大同府)	1881–1830	1
Songtao Ting (松桃厅)	Youyang Zhou (酉阳州)	1881–1830	1
Ningxia Fu (宁夏府)	Guisuiliuting (归绥六厅)	1881–1830	1
Ganzhou Fu (赣州府)	Yanping Fu (延平府)	1881–1830	1

Notes: This table lists the prefecture pairs identified as "outlier pairs" in the main text. These pairs exhibit a migrant share of 1, typically resulting from the small number of total out-migrants recorded in *XKTB* from the corresponding origin prefectures. To mitigate the potential influence of these outliers, in some estimations, we exclude all prefecture pairs in the corresponding period whose origin prefectures are listed in the first column. The Chinese names of the prefectures are provided in parentheses.

Table F.9 Estimating province boundary effects using province-pair dataset.

	Dependent variable: Migrant share among all individuals from the origin			
	(1)	(2)	(3)	(4)
Crossing *xiaqu* boundary	-1.271***	-1.266***	-1.263***	-1.261***
	(0.379)	(0.377)	(0.376)	(0.374)
Crossing province boundary	2.731***	0.093	-0.700	-1.164***
	(0.967)	(0.568)	(0.461)	(0.405)
Logarithm of distance × $Period_{1760s}$	-1.113***	-1.090***	-1.077***	-1.066***
	(0.202)	(0.208)	(0.212)	(0.215)
Logarithm of distance × $Period_{1820s}$	-1.128***	-1.113***	-1.104***	-1.096***
	(0.177)	(0.177)	(0.179)	(0.183)
Logarithm of distance × $Period_{1880s}$	-1.200***	-1.241***	-1.268***	-1.290***
	(0.194)	(0.201)	(0.207)	(0.213)
Migration dist. setting for local pairs	1 kilometers	10 kilometers	20 kilometers	30 kilometers
Origin FEs × Period FEs	Yes	Yes	Yes	Yes
Destination FEs × Period FEs	Yes	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes	Yes
# of individuals in *XKTB*	125,039	125,039	125,039	125,039
# of province-pair observations	1,587	1,587	1,587	1,587

Notes: This table presents the results from estimating Equation (13) with an additional variable indicating crossing province boundaries, using the province-pair datasets and the PPML estimator. All columns include constant terms, which are not listed in the table. Since individuals remaining within their local province include some inter-prefecture migrants, the average cost due to physical distance is not exactly zero. To assess the sensitivity of the estimates to the distance specification, we manually assign various migration distances for local pairs, ranging from 1 kilometer in Column (1) to 30 kilometers in Column (4). Robust standard errors, with two-way clustering at the origin and destination provinces, are shown in parentheses.
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.10 Estimating *xiaqu* boundary effects using province-pair dataset.

	Dependent variable: Migrant share among all individuals from the origin				
	1760s	1820s	1880s	All three periods	
	(1)	(2)	(3)	(4)	(5)
Crossing *xiaqu* boundary	-0.799***	-1.131***	-2.052***	-1.764***	-1.514***
	(0.296)	(0.280)	(0.500)	(0.434)	(0.341)
Logarithm of distance	-0.721***	-0.703***	-0.670***	-0.624***	
	(0.044)	(0.050)	(0.100)	(0.101)	
Clan difference	-4.249***	-5.421***	-2.687	-3.797	
	(1.340)	(1.568)	(2.706)	(2.708)	
Constant	-0.006	0.014	0.219***	0.036*	0.082***
	(0.007)	(0.012)	(0.040)	(0.019)	(0.016)
Origin fixed effects	Yes	Yes	Yes	Yes	No
Destination fixed effects	Yes	Yes	Yes	Yes	No
Period fixed effects	No	No	No	Yes	No
Origin FEs × Period FEs	No	No	No	No	Yes
Destination FEs × Period FEs	No	No	No	No	Yes
Controls × Period FEs	No	No	No	No	Yes
# of individuals in *XKTB*	45,046	50,255	29,738	125,039	125,039
# of province-pair observations	529	529	529	1,587	1,587

Notes: All estimations use the PPML estimator. Robust standard errors, with two-way clustering at the origin and destination provinces, are shown in parentheses.
***p < 0.01; **p < 0.05; *p < 0.1.

Table F.11 Robustness checks using province-pair dataset.

	Drop long-dist. migrants	Unweighted share	# of migrants	Polynomial of distance
	(1)	(2)	(3)	(4)
Crossing *xiaqu* boundary	-0.773***	-1.483***	-0.809***	-0.696***
	(0.216)	(0.273)	(0.227)	(0.262)
Origin FEs × Period FEs	Yes	Yes	Yes	Yes
Dest. FEs × Period FEs	Yes	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes	Yes
Polynomial of dist. × Period FEs	No	No	No	Yes
# of individuals in *XKTB*	124,566	125,039	125,039	125,039
# of province-pair obs.	714	1,587	1,587	1,587

Notes: This table presents robustness checks using the province-pair dataset, replicating the tests conducted with the prefecture-pair dataset. Column (1) excludes province pairs with migration distances exceeding 1,100 kilometers to reduce the influence of long-distance migrations. Column (2) calculates the migrant share using unweighted migrant numbers to ensure the results are not driven by reweighting. Column (3) uses the number of reweighted migrants as the dependent variable, instead of migrant shares. Column (4) replaces the logarithm of migration distance with a sixth-order polynomial of migration distance. All estimations use the PPML estimator and include constant terms that are not listed in the table. Robust standard errors, with two-way clustering at the origin and destination provinces, are shown in parentheses.
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.12 Heterogeneous *xiaqu* boundary effects.

	Dependent variable: Migrant share among all individuals from the origin
Crossing *xiaqu* boundary × Dongsansheng	-1.084*
	(0.582)
Crossing *xiaqu* boundary × Liangguang	-0.088
	(0.447)
Crossing *xiaqu* boundary × Liangjiang	-0.931**
	(0.452)
Crossing *xiaqu* boundary × Yungui	-1.334**
	(0.559)
Crossing *xiaqu* boundary × Neimenggu	4.781***
	(1.060)
Crossing *xiaqu* boundary × Sichuan	-0.667**
	(0.322)
Crossing *xiaqu* boundary × Shandong	-2.054***
	(0.360)
Crossing *xiaqu* boundary × Shanxi	-1.253***
	(0.228)
Crossing *xiaqu* boundary × Henan	-1.868***
	(0.413)
Crossing *xiaqu* boundary × Huguang	-0.915***
	(0.197)
Crossing *xiaqu* boundary × Zhili	-0.827**
	(0.388)
Crossing *xiaqu* boundary × Minzhe	-0.602
	(0.547)
Crossing *xiaqu* boundary × Shangan	-3.847***
	(1.244)
Origin FEs × Period FEs	Yes
Destination FEs × Period FEs	Yes
Controls × Period FEs	Yes
# of individuals in *XKTB*	125,039
# of province-pair observations	1,587

Notes: The estimation uses the PPML estimator and includes constant terms that are not listed in the table. Robust standard errors, with two-way clustering at the origin and destination provinces, are shown in parentheses.
***p < 0.01; **p < 0.05; *p < 0.1.

Table F.13 Heterogeneous *xiaqu* boundary effects, by period.

	Dependent variable: Migrant share among all individuals from the origin		
	1761–1770	1821–1880	1881–1890
Crossing *xiaqu* boundary × Dongsansheng	-1.561***	-0.570	-1.173
	(0.524)	(0.864)	(0.784)
Crossing *xiaqu* boundary × Liangguang	0.548	-0.243	-0.361
	(0.482)	(0.390)	(0.934)
Crossing *xiaqu* boundary × Liangjiang	-1.149***	-0.500	-0.687
	(0.425)	(0.790)	(0.651)
Crossing *xiaqu* boundary × Yungui	0.155	-2.393***	-2.535**
	(0.482)	(0.466)	(1.106)
Crossing *xiaqu* boundary × Neimenggu	3.342*		8.126***
	(1.943)		(1.336)
Crossing *xiaqu* boundary × Sichuan	-0.328	-0.596**	-0.830
	(0.329)	(0.261)	(0.608)
Crossing *xiaqu* boundary × Shandong	-1.587***	-2.231***	-2.198***
	(0.437)	(0.314)	(0.491)
Crossing *xiaqu* boundary × Shanxi	-0.998***	-0.964***	-1.514***
	(0.206)	(0.352)	(0.420)
Crossing *xiaqu* boundary × Henan	-1.584***	-2.169***	-1.596***
	(0.494)	(0.435)	(0.486)
Crossing *xiaqu* boundary × Huguang	-0.596	-1.502***	-0.120
	(0.389)	(0.401)	(0.456)
Crossing *xiaqu* boundary × Zhili	0.326	-0.998***	-1.672***
	(0.561)	(0.317)	(0.468)
Crossing *xiaqu* boundary × Minzhe	-1.009	-0.215	-0.357
	(0.825)	(0.796)	(0.510)
Crossing *xiaqu* boundary × Shangan	-1.716***	-1.450*	-5.800***
	(0.568)	(0.776)	(1.508)
Origin FEs × Period FEs	Yes	Yes	Yes
Destination FEs × Period FEs	Yes	Yes	Yes
Controls × Period FEs	Yes	Yes	Yes
# of individuals in *XKTB*	45,046	50,237	29,738
# of province-pair observations	529	506	529

Notes: The estimation uses the PPML estimator and includes constant terms that are not listed in the table. Robust standard errors, with two-way clustering at the origin and destination provinces, are shown in parentheses.
***p < 0.01; **p < 0.05; *p < 0.1.

Table F.14 Relative real income recovered from destination fixed effects, by province.

	By period		
	1761–1770	1821–1880	1881–1890
Yunnan (云南)	1.189	1.484	1.393
Sichuan (四川)	1.734	1.034	0.922
Anhui (安徽)	0.998	1.220	1.224
Shandong (山东)	0.952	0.911	0.847
Shanxi (山西)	0.693	0.780	0.847
Guangdong (广东)	1.054	1.125	0.983
Guangxi (广西)	1.160	1.385	1.247
Jiangsu (江苏)	1.248	1.439	1.419
Jiangxi (江西)	0.864	1.958	0.907
Henan (河南)	1.070	1.220	1.061
Zhejiang (浙江)	1.229	1.307	1.459
Hubei (湖北)	1.113	1.133	0.961
Hunan (湖南)	0.793	0.906	0.870
Gansu (甘肃)	1.255	1.607	1.022
Shengjing (盛京)	0.876	0.851	0.773
Zhili (直隶)	1.499	1.202	1.207
Fujian (福建)	0.793	1.038	1.755
Guizhou (贵州)	0.884	0.817	1.207
Shaanxi (陕西)	0.906	1.118	2.462

Notes: This table reports the relative real income compared to the national average by period, recovered using Equation (14) with destination fixed effects and a value of 2.5 for κ. Neimenggu (内蒙古), Jilin (吉林), Xinjiang (新疆), and Heilongjiang (黑龙江) are excluded from the table due to insufficient in-migrant samples to provide reliable estimates.

Table F.15 The fit of the Zipf's law, country level.

	Dependent variable: Logarithm of rank−0.5					
	1761–1770		1821–1830		1881–1890	
	(1)	(2)	(3)	(4)	(5)	(6)
Logarithm of population	-0.653***		-0.615***		-0.688***	
	(0.043)		(0.052)		(0.065)	
Logarithm of population density		-0.400***		-0.402***		-0.528***
		(0.038)		(0.038)		(0.048)
Constant	13.372***	8.024***	12.988***	8.127***	13.980***	9.230***
	(0.564)	(0.314)	(0.693)	(0.321)	(0.879)	(0.409)
R^2	0.657	0.504	0.619	0.499	0.657	0.559
# of observations	296	296	298	298	292	292

Notes: This table presents the fit of Zipf's law at the national level. Ranks in Columns (1), (3), and (5) are based on total population, while ranks in Columns (2), (4), and (6) are based on population density. Robust standard errors are provided in parentheses.
***p < 0.01; **p < 0.05; *p < 0.1.

Table F.16 The fit of the Zipf's law, *xiaqu* level.

	Dependent variable: Logarithm of rank−0.5					
	1761–1770		1821–1830		1881–1890	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Dongsansheng						
Logarithm of population	-0.925***		-0.537***		-0.394***	
	(0.101)		(0.074)		(0.089)	
Logarithm of population density		-0.478***		-0.491***		-0.513***
		(0.080)		(0.077)		(0.089)
R^2	0.919	0.895	0.867	0.901	0.689	0.877
# of observations	11	11	13	13	13	13
Panel B. Liangguang						
Logarithm of population	-0.790***		-0.779***		-0.767***	
	(0.187)		(0.186)		(0.184)	
Logarithm of population density		-1.537***		-1.540***		-1.519***
		(0.189)		(0.216)		(0.254)
R^2	0.695	0.905	0.680	0.883	0.665	0.849
# of observations	27	27	27	27	27	27
Panel C. Liangjiang						
Logarithm of population	-1.159***		-1.176***		-0.948***	
	(0.165)		(0.162)		(0.181)	
Logarithm of population density		-1.476***		-1.452***		-1.206***
		(0.148)		(0.149)		(0.137)
R^2	0.747	0.901	0.767	0.893	0.769	0.903
# of observations	39	39	39	39	39	39
Panel D. Yungui						
Logarithm of population	-1.054***		-1.045***		-1.056***	
	(0.155)		(0.155)		(0.170)	
Logarithm of population density		-1.405***		-1.443***		-1.444***
		(0.165)		(0.171)		(0.216)
R^2	0.740	0.845	0.731	0.842	0.739	0.721
# of observations	38	38	38	38	38	38
Panel E. Sichuan						
Logarithm of population	-0.588***		-0.554***		-0.502***	
	(0.121)		(0.118)		(0.102)	
Logarithm of population density		-0.485***		-0.456***		-0.429***
		(0.105)		(0.097)		(0.086)
R^2	0.648	0.535	0.625	0.516	0.602	0.507
# of observations	26	26	26	26	26	26
Panel F. Shandong						
Logarithm of population	-1.410***		-1.413***		-1.239***	
	(0.395)		(0.401)		(0.351)	
Logarithm of population density		-3.288***		-3.335***		-4.646***
		(0.233)		(0.228)		(0.378)
R^2	0.581	0.954	0.575	0.951	0.557	0.950
# of observations	12	12	12	12	12	12
Panel G. Shanxi						
Logarithm of population	-0.979***		-0.963***		-0.859***	
	(0.166)		(0.166)		(0.172)	
Logarithm of population density		-0.913***		-0.892***		-0.898***
		(0.265)		(0.254)		(0.233)
R^2	0.783	0.694	0.777	0.711	0.677	0.552
# of observations	20	20	20	20	20	20
Panel H. Henan						
Logarithm of population	-1.353***		-1.366***		-1.213***	
	(0.362)		(0.368)		(0.321)	
Logarithm of population density		-1.658***		-1.655***		-1.420**
		(0.514)		(0.510)		(0.581)
R^2	0.725	0.669	0.729	0.670	0.785	0.523
# of observations	13	13	13	13	13	13

Table F.16 (continued) The fit of the Zipf's law, *xiaqu* level.

	Dependent variable: Logarithm of rank−0.5					
	1761–1770		1821–1830		1881–1890	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel I. Huguang						
Logarithm of population	-0.557***		-0.574***		-0.632***	
	(0.128)		(0.131)		(0.148)	
Logarithm of population density		-1.638***		-1.708***		-2.160***
		(0.202)		(0.203)		(0.289)
R^2	0.541	0.867	0.560	0.878	0.575	0.828
# of observations	28	28	28	28	28	28
Panel J. Zhili						
Logarithm of population	-0.959***		-1.001***		-0.978***	
	(0.315)		(0.315)		(0.292)	
Logarithm of population density		-0.420***		-0.436***		-0.450**
		(0.135)		(0.146)		(0.162)
R^2	0.616	0.360	0.702	0.347	0.687	0.319
# of observations	18	18	18	18	18	18
Panel K. Minzhe						
Logarithm of population	-1.149***		-1.118***		-1.050***	
	(0.223)		(0.219)		(0.205)	
Logarithm of population density		-1.035***		-1.079***		-1.081***
		(0.160)		(0.138)		(0.175)
R^2	0.727	0.837	0.715	0.866	0.682	0.765
# of observations	23	23	23	23	23	23
Panel L. Shangan						
Logarithm of population	-0.795***		-0.886***		-0.909***	
	(0.141)		(0.191)		(0.191)	
Logarithm of population density		-0.558**		-0.558**		-0.550**
		(0.207)		(0.261)		(0.199)
R^2	0.757	0.527	0.735	0.450	0.731	0.544
# of observations	25	25	25	25	25	25

Notes: This table reports the fits of Zipf's law by *xiaqu*. The estimations include constant terms that are not listed in the table. Ranks in Columns (1), (3), and (5) are based on total population, while ranks in Columns (2), (4), and (6) are based on population density. Robust standard errors are shown in parentheses. Neimenggu, Shengjing, Heilongjiang, and Qinghai are not included in this table as there are too few effective prefectures for the estimation.

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table F.17 The fit of the Zipf's law, provincial level.

	Dependent variable: Logarithm of rank−0.5					
	1761–1770		1821–1830		1881–1890	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Yunnan						
Logarithm of population	-1.109***		-1.093***		-1.070***	
	(0.240)		(0.242)		(0.274)	
Logarithm of population density		-1.350***		-1.381***		-1.736***
		(0.121)		(0.126)		(0.238)
R^2	0.766	0.965	0.738	0.957	0.695	0.870
# of observations	22	22	22	22	22	22
Panel B. Sichuan						
Logarithm of population	-0.588***		-0.554***		-0.502***	
	(0.121)		(0.118)		(0.102)	
Logarithm of population density		-0.485***		-0.456***		-0.429***
		(0.105)		(0.097)		(0.086)
R^2	0.648	0.535	0.625	0.516	0.602	0.507
# of observations	26	26	26	26	26	26
Panel C. Anhui						
Logarithm of population	-0.898***		-0.908***		-0.694***	
	(0.213)		(0.214)		(0.186)	
Logarithm of population density		-2.349***		-2.333***		-1.088***
		(0.300)		(0.297)		(0.297)
R^2	0.676	0.928	0.679	0.931	0.765	0.667
# of observations	13	13	13	13	13	13
Panel D. Shandong						
Logarithm of population	-1.410***		-1.413***		-1.239***	
	(0.395)		(0.401)		(0.351)	
Logarithm of population density		-3.288***		-3.335***		-4.646***
		(0.233)		(0.228)		(0.378)
R^2	0.581	0.954	0.575	0.951	0.557	0.950
# of observations	12	12	12	12	12	12
Panel E. Shanxi						
Logarithm of population	-0.979***		-0.963***		-0.859***	
	(0.166)		(0.166)		(0.172)	
Logarithm of population density		-0.913***		-0.892***		-0.898***
		(0.265)		(0.254)		(0.233)
R^2	0.783	0.694	0.777	0.711	0.677	0.552
# of observations	20	20	20	20	20	20
Panel F. Guangdong						
Logarithm of population	-0.580***		-0.580***		-0.583***	
	(0.146)		(0.145)		(0.145)	
Logarithm of population density		-1.481***		-1.508***		-1.540***
		(0.270)		(0.278)		(0.289)
R^2	0.682	0.868	0.681	0.866	0.682	0.860
# of observations	15	15	15	15	15	15
Panel G. Guangxi						
Logarithm of population	-1.741***		-1.566***		-1.355***	
	(0.508)		(0.453)		(0.393)	
Logarithm of population density		-2.176***		-1.940***		-1.681**
		(0.601)		(0.584)		(0.547)
R^2	0.530	0.799	0.495	0.766	0.450	0.733
# of observations	12	12	12	12	12	12
Panel H. Jiangsu						
Logarithm of population	-1.187***		-1.218***		-1.369***	
	(0.340)		(0.332)		(0.257)	
Logarithm of population density		-1.194***		-1.122***		-1.498***
		(0.328)		(0.303)		(0.222)
R^2	0.693	0.621	0.731	0.586	0.854	0.897
# of observations	12	12	12	12	12	12

Table F.17 (continued) The fit of the Zipf's law, provincial level.

	Dependent variable: Logarithm of rank−0.5					
	1761–1770		1821–1830		1881–1890	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel I. Jiangxi						
Logarithm of population	-1.714***		-1.734***		-1.614***	
	(0.232)		(0.226)		(0.265)	
Logarithm of population density		-2.678***		-2.820***		-4.829***
		(0.485)		(0.504)		(1.141)
R^2	0.920	0.815	0.927	0.828	0.854	0.695
# of observations	14	14	14	14	14	14
Panel J. Henan						
Logarithm of population	-1.353***		-1.366***		-1.213***	
	(0.362)		(0.368)		(0.321)	
Logarithm of population density		-1.658***		-1.655***		-1.420**
		(0.514)		(0.510)		(0.581)
R^2	0.725	0.669	0.729	0.670	0.785	0.523
# of observations	13	13	13	13	13	13
Panel K. Zhejiang						
Logarithm of population	-1.850***		-1.761***		-1.409***	
	(0.353)		(0.328)		(0.307)	
Logarithm of population density		-1.101***		-1.097***		-1.214***
		(0.317)		(0.304)		(0.285)
R^2	0.828	0.770	0.810	0.771	0.775	0.732
# of observations	11	11	11	11	11	11
Panel L. Hubei						
Logarithm of population	-1.231***		-1.243***		-1.680***	
	(0.270)		(0.277)		(0.465)	
Logarithm of population density		-1.266***		-1.298***		-1.786***
		(0.228)		(0.228)		(0.338)
R^2	0.824	0.862	0.819	0.866	0.764	0.842
# of observations	11	11	11	11	11	11
Panel M. Hunan						
Logarithm of population	-0.500***		-0.519***		-0.581***	
	(0.120)		(0.122)		(0.127)	
Logarithm of population density		-1.837***		-1.962***		-2.318***
		(0.375)		(0.395)		(0.416)
R^2	0.584	0.731	0.615	0.755	0.692	0.773
# of observations	17	17	17	17	17	17
Panel N. Gansu						
Logarithm of population	-0.748**		-0.752**		-0.859***	
	(0.263)		(0.264)		(0.236)	
Logarithm of population density		-0.416*		-0.417*		-0.517**
		(0.197)		(0.198)		(0.224)
R^2	0.663	0.399	0.665	0.400	0.737	0.576
# of observations	13	13	13	13	13	13
Panel O. Zhili						
Logarithm of population	-0.959***		-1.001***		-0.978***	
	(0.315)		(0.315)		(0.292)	
Logarithm of population density		-0.420***		-0.436***		-0.450**
		(0.135)		(0.146)		(0.162)
R^2	0.616	0.360	0.702	0.347	0.687	0.319
# of observations	18	18	18	18	18	18
Panel P. Fujian						
Logarithm of population	-1.171***		-1.140***		-0.929***	
	(0.215)		(0.215)		(0.204)	
Logarithm of population density		-1.284***		-1.480***		-1.116***
		(0.267)		(0.226)		(0.184)
R^2	0.864	0.849	0.843	0.912	0.755	0.885
# of observations	12	12	12	12	12	12

Table F.17 (continued) The fit of the Zipf's law, provincial level.

	Dependent variable: Logarithm of rank−0.5					
	1761–1770		1821–1830		1881–1890	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel Q. Guizhou						
Logarithm of population	-0.930***		-0.933***		-0.957***	
	(0.173)		(0.168)		(0.167)	
Logarithm of population density		-2.545***		-2.888***		-3.035***
		(0.527)		(0.523)		(0.568)
R^2	0.730	0.753	0.743	0.800	0.785	0.823
# of observations	16	16	16	16	16	16
Panel R. Shaanxi						
Logarithm of population	-0.886***		-1.009***		-0.989***	
	(0.136)		(0.173)		(0.205)	
Logarithm of population density		-0.745***		-1.208***		-1.117***
		(0.172)		(0.238)		(0.310)
R^2	0.904	0.758	0.856	0.821	0.777	0.661
# of observations	12	12	12	12	12	12

Notes: This table reports the fits of Zipf's law by province. The estimations include constant terms that are not listed in the table. Ranks in Columns (1), (3), and (5) are based on total population, while ranks in Columns (2), (4), and (6) are based on population density. Robust standard errors are shown in parentheses. Neimenggu, Jilin, Shengjing, Xizang, Qinghai, and Heilongjiang are not included in this table as there are too few effective prefectures for the estimation.
$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Reference

Bailey, Mark. "Servile Migration and Gender in Late Medieval England: The Evidence of Manorial Court Rolls." *Past & Present* 261, no. 1 (2023): 47–85. https://doi.org/10.1093/pastj/gtac015

Barjamovic, Gojko, Thomas Chaney, Kerem Coşar, and Ali Hortaçsu. "Trade, Merchants, and the Lost Cities of the Bronze Age." *The Quarterly Journal of Economics* 134, no. 3 (2019): 1455–1503. https://doi.org/10.1093/qje/qjz009

Cao, Shuji. *Zhongguo Renkou Shi: Qing Shiqi (A History of the Chinese Population: The Qing Dynasty)*. Shanghai: Fudan University Press, 2001.

Chen, Jiafeng, and Jonathan Roth. "Logs with Zeros? Some Problems and Solutions." *The Quarterly Journal of Economics* 139, no. 2 (2024): 891–936. https://doi.org/10.1093/qje/qjad054

Chen, Zhiwu, Kaixiang Peng, and Lijun Zhu. "Social-economic change and its impact on violence: Homicide history of Qing China." *Explorations in Economic History* 63, (2017): 8–25. https://doi.org/10.1016/j.eeh.2016.12.001

Clark, Peter. "Migration in England during the Late Seventeenth and Early Eighteenth Centuries." *Past & Present* 83, no. 1 (1979): 57–90. https://doi.org/10.1093/past/83.1.57

Dingel, Jonathan I., and Felix Tintelnot. "Spatial Economics for Granular Settings." Working Paper Series Working Paper 2020. https://doi.org/10.3386/w27287

Lal, Apoorva, Mackenzie Lockhart, Yiqing Xu, and Ziwen Zu. "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies." *Political Analysis* (2024): 1–20. https://doi.org/10.1017/pan.2024.2

Lee, James Z., and Cameron D. Campbell. "China Multi-Generational Panel Dataset, Liaoning (CMGPD-LN), 1749–1909." 2016. https://doi.org/10.3886/ICPSR27063.v10

Liu, Ziang. "Wages, labour markets, and living standards in China, 1530–1840." *Explorations in Economic History* 92, (2024): 101569. https://doi.org/10.1016/j.eeh.2023.101569

Silva, J. M. C. Santos, and Silvana Tenreyro. "The Log of Gravity." *The Review of Economics and Statistics* 88, no. 4 (2006): 641–658. https://doi.org/10.1162/rest.88.4.641

Sotelo, Sebastian. "Practical Aspects of Implementing the Multinomial PML Estimator." 2019.

Tombe, Trevor, and Xiaodong Zhu. "Trade, Migration, and Productivity: A Quantitative Analysis of China." *The American Economic Review* 109, no. 5 (2019): 1843–1872. https://doi.org/10.1257/aer.20150811

Yang, Cheng. "A new estimate of Chinese male occupational structure during 1734–1898 by sector, sub-sector pattern, and region." *The Economic History Review* 75, no. 4 (2022): 1270–1313. https://doi.org/10.1111/ehr.13157