

The Journal of International Trade & Economic Development

An International and Comparative Review

ISSN: 0963-8199 (Print) 1469-9559 (Online) Journal homepage: www.tandfonline.com/journals/rjte20

Job prospects and labour mobility in China

Huaxin Wang-Lu & Octasiano Miguel Valerio Mendoza

To cite this article: Huaxin Wang-Lu & Octasiano Miguel Valerio Mendoza (2023) Job prospects and labour mobility in China, The Journal of International Trade & Economic Development, 32:7, 991-1034, DOI: [10.1080/09638199.2022.2157463](https://doi.org/10.1080/09638199.2022.2157463)

To link to this article: <https://doi.org/10.1080/09638199.2022.2157463>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 22 Dec 2022.



Submit your article to this journal [↗](#)



Article views: 4994



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 8 View citing articles [↗](#)

Job prospects and labour mobility in China

Huaxin Wang-Lu  and Octasiano Miguel Valerio Mendoza 

Department of Quantitative Methods, IQS School of Management, Universitat Ramon Llull, Barcelona, Spain

ABSTRACT

China's structural changes have brought new challenges to its regional employment structures, entailing labour redistribution. Until now, Chinese migration research with a forward-looking perspective and on bilateral longitudinal determinants at the prefecture city level is almost non-existent. This paper investigates the effects of job prospects on individual migration decisions across prefecture boundaries. To this end, we created proxy variables for wage and employment prospects, introduced reference-dependence to a dynamic discrete choice model, and estimated corresponding empirical specifications with a unique quasi-panel of 66,427 individuals from 283 cities during 1997–2017. To address multilateral resistance to migration resulting from the future attractiveness, we exploited various monadic and dyadic fixed effects. Multilevel logit models and two-step system GMM estimation were adopted for the robustness check. Our primary findings are that a 10% increase in the ratio of sector-based employment prospects in cities of destination to cities of origin raises the probability of migration by 1.281–2.185 percentage points, and the effects tend to be stronger when the scale of the ratio is larger. Having a family migration network causes an increase of approximately 6 percentage points in migratory probabilities. Further, labour migrants are more likely to be male, unmarried, younger, or more educated. Our results suggest that the ongoing industrial reform in China influences labour mobility between cities, providing important insights for regional policymakers to prevent brain drain and to attract relevant talent.

KEYWORDS Internal migration; labour mobility; expectations; reference-dependence; China

JEL CLASSIFICATIONS J61, D90, R11, O14, O15

ARTICLE HISTORY Received 21 July 2022; Accepted 5 December 2022

1. Introduction

Countries often undergo a drastic change in their employment structures at times of industrial transformation, entailing labour redistribution. China is a remarkable case for study under this background. In terms of China Statistical Yearbooks, since the beginning of its reforms and opening-up in late 1978, the GDP-based ratio of the primary,

CONTACT Huaxin Wang-Lu  huaxin.wanglu@iqs.url.edu; huaxin.wanglu@gmail.com  Department of Quantitative Methods, IQS School of Management, Universitat Ramon Llull, Via Augusta, 390, Barcelona 08017, Spain

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

secondary and tertiary sector changed from 31:47:22 in 1979 to 8:38:54 in 2020. This indicates that the tertiary sector almost trebled its contribution to national GDP during this period. In recent years, after being the ‘world’s factory’ for decades, the central government launched the ‘Made in China 2025’ and ‘Dual Circulation’ strategies to promote upgrading from a labour-intensive, export-oriented manufacturing economy to a service and consumption-driven one.

At the same time, China unveiled a proposed revision to the law on vocational education, announcing in 2020 the ‘Vocational Education Quality Improvement Action Plan’ to fill gaps in skilled technicians and to differentiate skillsets across college graduates. On the demand side, regions at different levels of development require different types of skills at different levels of demand. For instance, inland provinces need workers with plant-based skills in response to the relocation of numerous factories from eastern areas taking advantage of lower labour costs (Qu, Cai, and Zhang 2013), whereas the Yangtze River Delta is dedicated to attracting high-tech and managerial professionals for highly-developed manufacturing and service industries (Wang et al. 2020). On the supply side, the job prospects for individuals with diverse skills and profiles differ across regions.

In general, the employment situation of workers with low and average skills is dim, because of the widespread use of automated technology, although the service sector is creating new low-tech jobs (Li et al. 2020). More recently, computerisation has added to jobless growth.¹ Frey and Osborne (2017) assert that the majority of routine jobs in manufacturing and a range of sub-sectors within the tertiary sector, such as finance, logistics and administrative support, are at high risk of being replaced by artificial intelligence (AI) technology, a potential concern for China. Zhou et al. (2020) even suggest that AI technology will replace approximately 278 million jobs in China by 2049. While the creation of non-routine jobs was stagnant from 1990 to 2015, with more than 50% of employment routinised in 2015 (Ge, Sun, and Zhao 2021).²

The extent of all these (expected) impacts varies across sectors and regions. It is inevitable that employment prospects will be subject to change in regional employment structures, stimulating relocation, a plausible reaction to worsening prospects in particular regions. These impacts could go beyond concerns about job opportunities into job mobility and/or security. In other words, labour market conditions tend to be inconsistent with economic growth, at least in the short term (e.g. Şahin, Tansel, and Berument 2015). These come to our core idea that an employment flourishing industry could provide better job prospects for individuals with migration intentions over those whose job growth shrinks or stagnates, *ceteris paribus*. As seen in Figure 1, GDP and wages grew smoothly in Beijing, Shanghai, Guangzhou, and Shenzhen, while employment levels in the four cities fluctuated. While on the other hand, cities with more increases in wages could have a ‘comparative advantage’ to further benefit job prospects of their workers. Thus, we also consider expected earnings as a useful complementarity to the assessment of job prospects of potential migrants. In sum, we divide job prospect into two components, that is, employment and wage prospects.

Then, the underlying question is how do potential migrants make their assessments? For simplicity, we use income as an example for elaboration. Suppose that all individuals in city A always earn 2000 Yuan per month. It will be difficult for them to make an assessment of economic conditions, assuming no inflation. At the same time, people in city B earn 2000 Yuan per month in 2011 but 1500 Yuan in 2012. The latter population is very likely to believe that the economy is getting worse, at least, more likely than

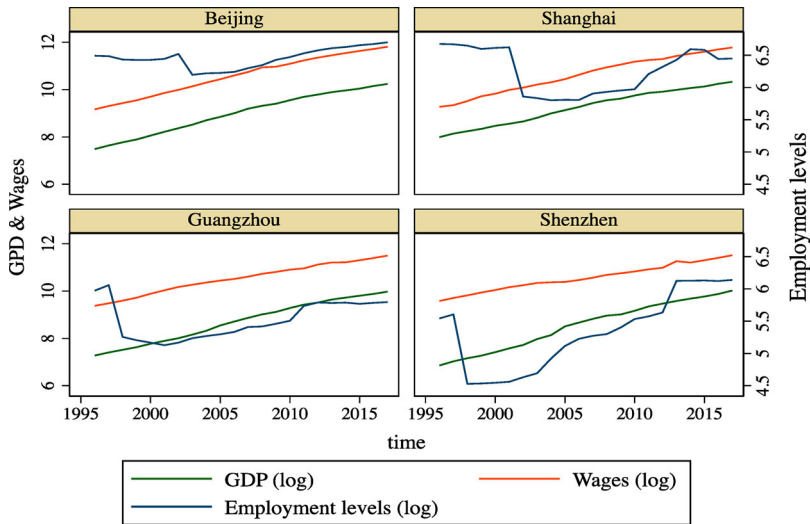


Figure 1. GDP, wages vs. employment levels in 'Bei-Shang-Guang-Shen'.

Author's elaboration using China Data Institute (2021).

are A residents. As prospect theory reveals (Kahneman and Tversky 1979),³ the deviation (1500–2000) from the 2011 income, a reference point closest to the 2012 income, generates a negative signal for evaluation and accordingly incurs a pain of loss to B inhabitants.

Let's consider another two cities C and D where people earn, respectively, 2000 and 3000 Yuan per month in 2011 but both 2500 Yuan in 2012. If B inhabitants with migration intentions are aware of the changes, although moving to either city C or D will lead to the same increases in their income, they will perceive additional gains from C 's positive deviation (2500–2000). This is because such an uptrend implies that the prospect in city C is better than in the others. Kahneman and Tversky highlighted that individuals would even reverse their preferences when identical outcomes are rephrased as gains or losses. As outlined in Figure 2, this example manifests why reference-dependence is meaningful in our context.

On the other hand, the feature of population redistribution has also been changing. Intra-provincial migration seems to have been more important than inter-provincial migration for both temporary and permanent migrants during this decade (Zhang and Zhao 2013; Meng 2020). Inter-provincial inequalities have often been associated with the prevalence of inter-provincial migration (e.g. Peng and Swider 2017). If this is the case, a co-movement of the popularity of intra-provincial migration and intra-provincial inequalities will arise. Hence, keeping a watchful eye on city-level migration is of greater importance to China than ever before.⁴

The objective of this paper is to examine the effects of job prospects on individual migration decisions across prefecture boundaries.⁵ Our contribution to the migration debate is fourfold. Firstly, we contributed to the considerable dearth in Chinese migration literature of visionary migration decisions concerned with optimising future outcomes.⁶ Though a handful of studies touch upon migrants' expectations, such as expected land reallocation (e.g. Ren et al. 2020), none considers multilateral resistance

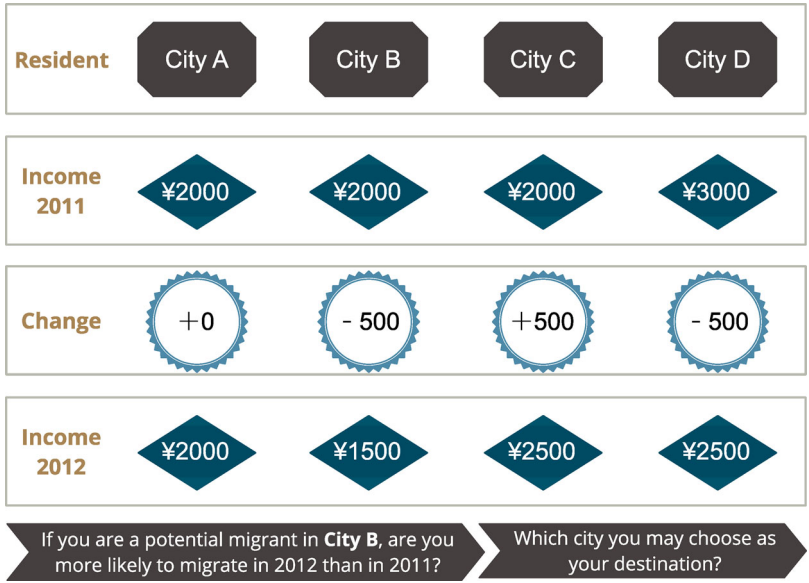


Figure 2. Example for understanding reference-dependence.

to migration resulting from the future attractiveness (Bertoli and Fernández-Huertas Moraga 2013).⁷ We understand that this is the first paper to illustrate migration decision-making of visionary labour migrants within China. Secondly, we extended the random utility maximisation (RUM) model of migration by synthesising the virtues of dynamic discrete choice modelling framework and reference-dependence to further illustrate the formation of individual expectations.⁸ To our knowledge, this is also the first time that linkages have been established between the RUM model of migration and prospect theory (Kahneman and Tversky 1979). Thirdly, we derived empirical specifications with theoretical micro-foundations. Lastly, we compiled a unique quasi-panel of 66,427 individuals moving from 283 cities to 279 cities during 1997–2017 and thus combined city-level bilateral variations with individual and household characteristics, a level of analysis not yet undertaken by the existing Chinese migration literature.⁹ Both the new migration patterns and the essential role of prefecture cities in managing Hukou registration signify the importance of understanding city-level push and pull factors.

2. Migration and expectations

2.1. Migration in China

Despite the surge of temporary migration since the late 1980s, this phenomenon was not studied empirically until much later, principally due to the lack of data. The earlier migration literature relied mainly on macro-data, typically the national population census or the 1% population sample survey, treating intra-provincial migrants as ‘non-movers’ (e.g. Poncet 2006). Although more micro-data became available in the late 2000s, e.g. the Longitudinal Survey on Rural-Urban Migration in China (RUMiC) initiated in 2006, and the China Migrants Dynamic Survey (CMDs) started in 2009, the vast majority of data options remain incompatible with bilateral or inter-city studies.¹⁰ More

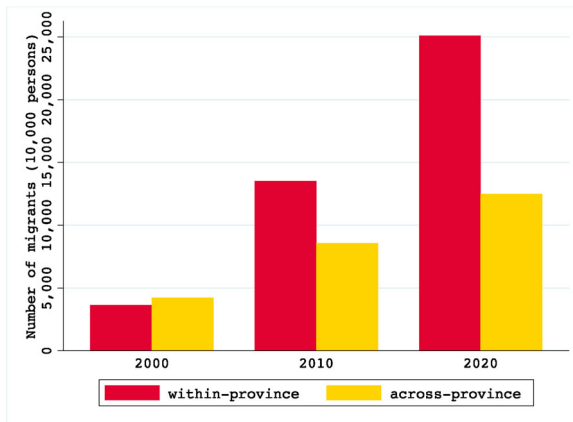


Figure 3. The number of intra-provincial migrants vs. inter-provincial migrants.
The 2000, 2010 and 2020 Chinese Censuses.

recently, micro-data have been popular for migration analysis as they can help alleviate reverse causality problems and allow researchers to control for individual heterogeneity of relevance. However, longitudinal micro-data are much harder to collect, resulting in only a few migration studies on individual decisions over time. Further, the popularity of the province-level research also stems from the remarkable increase in inter-provincial migrant populations since 1987 (Liang 2001), the predominant group until 2010, which was then outpaced by intra-provincial migration, as seen in Figure 3. Similar conclusions can be drawn from national statistics on the rural-to-urban subpopulation, the main focus of Chinese migration studies. By combining the 1990, 2000, and 2010 Chinese Censuses with the 1995, 2005, and 2015 1% population sample surveys, Su, Tesfazion, and Zhao (2018) show that intra-provincial rural migration flows have outnumbered inter-provincial flows since 2011 and empirically identify that the likelihood of moving within the province of origin increases for rural migrants who are older, more educated, female, single, or from poorer areas.

Determinants of migration are often conceptualised as push-and-pull factors associated with the sending and receiving locations, respectively. Economic reasons, such as income gaps, job opportunities, and land tenure insecurity, are often the most decisive factors in a substantial body of literature (e.g. Liu and Shen 2014). However, even though migrants usually move for monetary reasons, contributing factors vary between groups and over time. Demographically, researchers continually find that age, gender, marital status, education level, and family size can explain the probability of migration (e.g. He and Gober 2003; Mu et al. 2021). Spatial and social factors, such as geographic distance, urban amenities, public policies, and the Hukou registration barrier, also play a vital role in propelling or discouraging migration (e.g. Zhang et al. 2020). The academic debate on the weight of economic opportunities vs. amenities in migration has accordingly received more attention (e.g. Wang et al. 2020). Additionally, there is an increasing emphasis in the migration literature on environmental degradation, such as climate change and air pollution, caused mainly by industrialisation and over-population (e.g. Liu and Yu 2020).

In addition to these determinants, the impacts of migration on migrants' families, such as consumption patterns and left-behind children (e.g. Meng and Xue 2020; Li and Luo 2021), impacts on areas of destination, such as skill aggregation, wage premiums, and occupational upgrading (e.g. Chung, Zhang, and Partridge 2020; Zhao 2020), impacts on other potential migrants, such as networks and co-location (e.g. Foltz, Guo, and Yao 2020; Fu and Gabriel 2012), and impacts on the national economy if international movers are present, such as trade and foreign direct investment (Hatzigeorgiou and Lodefalk 2021), among others, have been widely discussed.

2.2. *Expectations in migration decision-making*

Individual expectations about future outcomes are demonstrated as a compelling driving force behind migration decisions. For instance, De Jong (2000) found that income expectancies of remaining in home communities vs. living in alternative locations, in addition to residential satisfaction, were key determinants of migration intentions for both men and women in Thailand. Baldwin (2001) extended the core-periphery model, illustrating that the prevailing assumption of myopic migrants holds when migration costs are high, while forward-looking expectations arise in scenarios where migration costs are relatively low. More recently, Baumann, Svec, and Sanzari (2015) developed a Harris-Todaro model and used US state-level data to show that unemployment rates *per se* did not affect migration, but rather that changes in residents' expectations of unemployment across regions induced migration. Likewise, Shrestha (2020) conducted a randomised field experiment to demonstrate that individual expectations changed by gaining information on earnings and mortality rates abroad influenced actual migration decisions of potential migrants in Nepal, particularly of inexperienced migrants.

Other studies of note are of two types. The first is grounded in discrete choice models initiated by McFadden (1974), which provides a theoretical basis for gravity models of migration.¹¹ Here, a canonical RUM model includes a deterministic and a stochastic component of utility and a time-specific migration cost. The distributional assumptions on the stochastic term determine the expected probability of selecting a destination. The deterministic term is typically modelled as a function of state variables, such as income, population, or temperature.¹² Bertoli, Brücker, and Fernández-Huertas Moraga (2016) expanded the RUM model to allow migrants to make sequential decisions, such as return to their origin cities or move to alternative locations after migration. They assumed that migrants who are neither myopic nor living in a frictionless world where migration costs are zero respond to the future attractiveness and accessibility of all locations in the choice set. By using two proxies for expectations, they found that economic prospects at origin significantly influenced the scale of migration to Germany. Beine, Bourgeon, and Bricogne (2019) also modified the RUM model but contended that agents' expectations of future outcomes were formed by the current level of economic activity and employment rates both within the country of origin and in a number of potential destination countries. Two proxies for signalling expectations about future employment probabilities at destination are found to have been influential in bilateral migration flows.

Another study type considers behavioural theories. Czaika (2015) developed the migration prospect theory grounded on Kahneman and Tversky's (1979) seminal work on prospect theory. Prospect theory suggests that the utility of an agent does not depend on an absolute outcome but on gains or losses gradually perceived relative to

the reference point, or, in other words, on deviations from the status quo. By applying fundamental principles of prospect theory, i.e. loss aversion, reference-dependence, risk preferences, and diminishing sensitivity, to the migration decision-making model, he demonstrated that migration flows responded more strongly to negative economic and unemployment prospects in home countries than to equal-sized positive prospects in Germany. Clark and Lisowski (2017) have also gained insights from prospect theory, emphasising the endowment effect whereby people place a higher use value on the object they own than its market value to explain residential moving or staying. Their findings showed that, for internal migration in Australia, the probability of staying increases with stronger risk aversion, is higher for owners than for renters and is higher still with longer duration at the current address.¹³

3. Theoretical and empirical framework

3.1. A reference-dependent migration model

When individuals are visionary, the location-specific utility that reflects the sequential nature of the decision-making problem defined by Bertoli, Brücker, and Fernández-Huertas Moraga (2016) based on dynamic discrete choice models (e.g. Kennan and Walker 2011) is as follows:

$$U_{ijk,t} = w_{kt} + \beta A_{kt+1}(I) - c_{jk,t} + \epsilon_{ijk,t} \quad (1)$$

where $U_{ijk,t}$ is the utility of an individual i who moves from city j to city k at time t . w_{kt} is the deterministic instantaneous component of utility gained by moving from city j to city k at a time t . $c_{jk,t}$ describes the cost of moving from j to k at t . $\epsilon_{ijk,t}$ is an individual stochastic and time-specific serially-uncorrelated component of utility. Both w_{kt} and $c_{jk,t}$ are known to the individual. The expected utility gained by moving from city j to city k at time t and optimally choosing the preferred location from $t + 1$ onward is $A_{kt+1}(I)$. We add I here to denote that the location preference hinges on the industry to which job categories that individual i searches for belong. Jobs sought at destination are not necessarily identical to individual i 's previous work at origin. $\beta \in [0, 1)$ is the time discount factor of the expected utility. $\beta = 0$ represents the fact that potential migrants do not attach importance to future outcomes and as a result, make myopic decisions. It is also assumed that individual i chooses the preferred location after being aware of the stochastic component of utility at time t for all cities.

Beine, Bourgeon, and Bricongne (2019) express the same idea that future outcomes matter, but assume that individual expectations are formed by extracting information from current economic conditions. Their model can be written as:

$$U_{ijk,t} = w_{kt} + \ln(E[A_{kt}(I)]) - c_{jk,t} + \epsilon_{ijk,t} \quad (2)$$

Although the model is constructed differently, $E[A_{kt}(I)]$ is basically the same to individual i as the discounted value $\beta A_{kt+1}(I)$ at time t , but will be different at some time $t + g \geq t + 1$ if individual i moves away from k , because the expected instantaneous utility implies a permanent stay or move (Kennan and Walker 2011). Furthermore, equation (2) takes the log of the expected utility to express the non-linearity and constant relative risk aversion.

Czaika (2015) also takes into account the non-linearity of utility and risk attitudes via replacing the utility function over absolute outcomes with the value function over gains

and losses relative to a reference point, as does prospect theory. The reference-dependent migration value function derived from Kőszegi and Rabin (2009) can be written as:

$$V_{it} = V_{it}^k - V_{it}^j = M(\tilde{y}_{it}^k - \tilde{y}_{it}^j) + N(y_{it+1}^k, y_{it+1}^j | y_{it}^k, y_{it}^j) \quad (3)$$

$$\text{where } N(\cdot) = (y_{it+1}^k - y_{it}^k - y_{it+1}^j + y_{it}^j)^\alpha = (\Delta_{it+1}y^k - \Delta_{it+1}y^j)^\alpha \quad (4)$$

where V_{it} is the value of migration for individual i who moves from city j to city k at time t . $M(\cdot)$ is the regular component drawn from absolute outcomes in the origin city $\tilde{y}_{it}^j$ and the destination city $\tilde{y}_{it}^k$. In contrast, $N(\cdot)$ is the reference-dependent utility where present economic situations in the origin and destination city, i.e. y_{it}^k and y_{it}^j , respectively act as a reference point in adjusting present expectations about the future, i.e. y_{it+1}^k and y_{it+1}^j . The superscript α emphasises the non-linearity.¹⁴

Here, we fine-tune the expected continuation payoff $A_{kt+1}(I)$ in equation (1) to be reference-dependent, as follows:

$$A_{kt+1}(I) = y_{it+1}^k - y_{it}^k = \Delta_{it+1}y^k(I); \quad A_{jt+1}(I) = y_{it+1}^j - y_{it}^j = \Delta_{it+1}y^j(I) \quad (5)$$

In a one-time migration scenario, we can rewrite equation (5) as:

$$A_{kt+g}(I) = \frac{1}{\beta^g} E[A_{kt}(I)] = \frac{1}{\beta^g} E[A_{kt}(I) | r^g] \cdot E[r^g] = \frac{r^g}{\beta^g} \cdot \Delta_{it}y^k(I) \quad (6)$$

where $r \in (0, 1)$ denotes the uncertainty between present trends and future realisations. $t + g \geq t + 1$ is any future point in time from time $t + 1$ onward. $r^g \Delta_{it}y^k(I)$ indicates that the further the future, the greater the uncertainty. Similarly, $\beta^g A_{kt+g}(I)$ can be understood as the further the future, the less influential will it be to the present utility.

Nevertheless, the continuation payoff $\Delta_{it+1}y^k(I)$ derived from Bertoli, Brücker, and Fernández-Huertas Moraga (2016) entails a more intriguing intuition than the one-time approach that individuals can update their reference points after moving to k at time t and choose any alternative location q among the choice set D at time $t + 1$ in terms of their new reference points. If we assume that the stochastic component of utility follows an independent and identically distributed (i.i.d) Extreme Value Type-1 distribution (McFadden 1974) with zero mean where τ is the Euler constant, the recursive form of the expected utility conditional on residing in k at time $t + 1$, in terms of Small and Rosen (1981) and Kennan and Walker (2011), can be expressed as:¹⁵

$$\Delta_{it+1}y^k(I) = \tau + \ln \left(\sum_{q \in D} e^{w_{qt+1} - c_{kq,t+1} + \beta \Delta_{it+2}y^q(I)} \right) \quad (7)$$

Then, equation (1) can be correspondingly rewritten as:

$$U_{ijk,t} = w_{kt} + \beta \left(\tau + \ln \left(\sum_{q \in D} e^{w_{qt+1} - c_{kq,t+1} + \beta \Delta_{it+2}y^q(I)} \right) \right) - c_{jk,t} + \epsilon_{ijk,t} \quad (8)$$

As McFadden (1974) shows, the probability of migrating from city j to city k can be estimated as:

$$P_{ijk,t} = \Pr \left\{ U_{ijk,t} = \max_{q \in D} U_{ijq,t} \right\} = \frac{e^{U_{ijk,t}}}{\sum_{q \in D} e^{U_{ijq,t}}} \quad (9)$$

$$\ln \left(\frac{P_{ijk,t}}{P_{ijj,t}} \right) = w_{kt} - w_{jt} - c_{jk} + \beta \cdot [\Delta_{it+1}y^k(I) - \Delta_{it+1}y^j(I)]$$

where w_{jt} is the utility for individual i choosing to remain in city j at time t . The probability of k being chosen over j among the choice set D is equivalent to the probability of a binary choice of k over j if we assume the denominator is positive for all possible alternative choices.¹⁶

So far, we do not account for the non-linearity of utility as seen in equations (2) and (4). It is very plausible that values attached to the migration project are not marginally linear but subject to the scale of change in economic situations, particularly, of the reference point. We can extend equation (5) by taking the logarithm of $\Delta_{it+1}y^k(I)$ and $\Delta_{it+1}y^j(I)$. Thus, equation (9) can be rewritten as:

$$\ln \left(\frac{P_{ijk,t}}{P_{ijj,t}} \right) = w_{kt} - w_{jt} - c_{jk} + \beta \cdot \ln \frac{\Delta_{it+1}y^k(I)}{\Delta_{it+1}y^j(I)} \quad (10)$$

3.2. Econometric modelling and techniques

Our innovative predictor is the main proxy for job prospects in the city of origin and city of potential destination, respectively. We use the term ‘trending’ to highlight the fact that it captures upward and downward trends at the industry level. Macro surroundings often silently yet profoundly influence individual perceptions, and accordingly, trending signals derived here mirror the role of contextual evolution in forming expectations of all relevant individuals.

$$E_Trending_{ij,t} = \frac{E_{ij,t} - E_{ij,t-1}}{E_{ij,t-1}} - \frac{E_{ij,t-1} - E_{ij,t-2}}{E_{ij,t-2}} = GR_{ij,t} - GR_{ij,t-1} = \Delta_t GR_{ij} \quad (11)$$

$$E_Trending_{ik,t} = \frac{E_{ik,t} - E_{ik,t-1}}{E_{ik,t-1}} - \frac{E_{ik,t-1} - E_{ik,t-2}}{E_{ik,t-2}} = GR_{ik,t} - GR_{ik,t-1} = \Delta_t GR_{ik} \quad (12)$$

where the quantity of employment at time t in the sector of job categories that individual i looks for is $E_{ij,t}$ for the origin city j and $E_{ik,t}$ for the destination city k . In short, the trending indicator is the annual change in growth rates of industrial employment, i.e. $\Delta_t GR_{ij}$ and $\Delta_t GR_{ik}$. The beauty of this design is that positive growth does not necessarily lead to better employment prospects and vice versa. If we consider that $GR_{ij,t} = 1.5\%$ and $GR_{ij,t-1} = 1.8\%$, despite both being positive, the outcome is -0.3% signalling a slowdown. Likewise, for negative growth over two years, such as $GR_{ij,t} = -0.5\%$ and $GR_{ij,t-1} = -0.7\%$, $\Delta_t GR_{ij} = 0.2\%$ is still a positive value, because the scale of the present decline is narrower, implying that employment prospects stand a chance of getting better. In other words, individuals gain utility from staying or moving, not because of the growth *per se*, but due to whether progress is faster (better) or slower (worse) relative to the previous year at origin vs. destination.

The implication for this construction inspired by the migration prospect theory (Czaika 2015) can be linked to Baumann, Svec, and Sanzari's (2015) findings that unemployment affects migration only if it alters expectations, and the central concept of Clark and Lisowski (2017). In our context, it could be understood that $GR_{ij,t} = 1.5\%$ does not trigger migration since no expected gain or loss emerges if $GR_{ij,t-1}$ also $= 1.5\%$. Moreover, Clark and Lisowski (2017) elaborate what prospect theory could offer for understanding why the majority of people prefer staying, because, in terms of loss aversion, people do not necessarily choose the highest expected utility gained via migration if they are more concerned about losing what they have. This offers insights into understanding our indicator from another angle. If we consider an industry with consistent growth at both origin and destination, to compete against the status quo bias ingrained in resident workers, deviations from two reference points are critical to add weight to moving and, consequently, to impel decision-making.¹⁷

Thus, we can formulate the distance of expected utility between the destination and origin city:

$$Distance(E_Trending_{ijk,t}) = \Delta_t GR_{ik} - \Delta_t GR_{ij} \quad (13)$$

The distance variable corresponding to equation (10) can be calculated as:

$$Distance^T(E_Trending_{ijk,t}) = \ln(\Delta_t GR_{ik}) - \ln(\Delta_t GR_{ij}) = \ln\left(\frac{\Delta_t GR_{ik}}{\Delta_t GR_{ij}}\right) \quad (14)$$

where $Distance^T$ is the transformed version of equation (13). The assumption is that holding the difference in employment prospects constant, potential migrants are more reluctant to move when their expectations of employment prospects in their origin cities are good, and the better the prospects, the less responsive potential migrants become (e.g. given $\Delta_t GR_{ik} - \Delta_t GR_{ij} = 1$, $\ln(2) - \ln(1) = 0.69$ vs. $\ln(10) - \ln(9) = 0.11$). Conversely, the worse the local prospects, the more susceptible potential migrants become.¹⁸

Likewise, we create a distance variable of reference-dependent wages as an additional proxy for job prospects:

$$Distance^T(Wage_{ijk,t}) = \ln(\Delta_t Wage_{ik}) - \ln(\Delta_t Wage_{ij}) = \ln\left(\frac{\Delta_t Wage_{ik}}{\Delta_t Wage_{ij}}\right) \quad (15)$$

where $\Delta_t Wage_{ik}$ and $\Delta_t Wage_{ij}$ are annual differences in average wages of staff and workers in city k and city j , respectively.¹⁹ It is noteworthy that both our trending and wage indicators are the empirical counterpart of $\Delta_{it+1}y^j(I)$ and $\Delta_{it+1}y^k(I)$.

As binary outcomes cannot be log-transformed, our fixed-effects specifications act as a linear probability model approximation to equation (10), which can be accurate in magnitude and sign for small parameter values in practical matters (McFadden 1974). The baseline to be estimated is as follows:

$$M_{ijk,t} = \alpha + \beta_1 X_{it} + \beta_2 Distance^T(Prospect_{ijk,t}) + \beta_3 Distance(Z_{ijk,t-1}) + \gamma_t + \epsilon_{ijk,t} \quad (16)$$

where the binary dependent variable (DV) $M_{ijk,t}$ equals 1 if individual i chooses to move from city j to city k at time t , and 0 otherwise. $Prospect_{ijk,t}$ could be either $E_Trending_{ijk,t}$ or $Wage_{ijk,t}$. X_{it} is a row vector of individual and household characteristics. $Distance(Z_{ijk,t-1}) = Z_{ik,t-1} - Z_{ij,t-1}$ are city-level control variables. Because these

covariates are state variables measured at the end of each year, we lag them by one period. γ_t are time-fixed effects and $\epsilon_{ijk,t}$ is an idiosyncratic error term.

In this paper, the city covariates refer to income level, usually measured through GDP per capita (Beine, Bertoli, and Fernández-Huertas Moraga 2016), population density in relation to traffic congestion and informal settlement and slums (Tan, Xu, and Zhang 2016), and urban amenities associated with two major aspects: the provision of healthcare (the number of hospital beds per person) and higher education resources (the number of higher educational institutions per 10,000 persons) (Czaika and Parsons 2017). To avoid multicollinearity, we replace GDP per capita with unemployment rates for estimating wage indicators. We also include the share of the tertiary sector and the number of enterprises above the designated size per 10,000 persons to account for gaps in regional commercialisation and business density, respectively.²⁰ Additionally, in some specifications, we add the China Hukou Registration Index (CHRI) developed by Zhang, Wang, and Lu (2019) for 120 Chinese cities to control for Hukou entry policies and migration costs.²¹ The individual and household covariates are: gender, marital status, Hukou type (rural or urban), self-evaluated health status, household income, age, years of schooling, and family migration network. Research suggests that the presence of networks lowers migration costs, increases the probability of migration and is more important for the mobility of low-educated migrants than for the higher-educated (e.g. Beine, Bourgeon, and Bricongne 2019). We construct it as a dummy, indicating if there are any pioneer migrants within families prior to the move of each individual.

We then add three fixed effects to equation (16) to alleviate endogeneity issues, as follows:

$$M_{ijk,t} = \alpha + \beta_1 X_{it} + \beta_2 \text{Distance}^T(\text{Prospect}_{ijk,t}) + \beta_3 \text{Distance}(Z_{ijk,t-1}) + \gamma_t + \gamma_j + \gamma_k + \gamma_s + \epsilon_{ijk,t} \quad (17)$$

where γ_j , γ_k and γ_s are origin, destination and industrial sector fixed effects. Adding them helps reduce unobservable time-invariant or very slowly varying push and pull factors of each origin, destination, and industrial sector, such as the Hukou registration policies in origin or destination cities, industry-specific policies, or demographic characteristics of the population in each city or industry. This is the traditional strategy to control for multilateral resistance to migration in cross-sectional studies (Mayda 2010). In such cases, multilateral resistance to migration occurs when a destination city has different levels of attractiveness to people from the same place of origin due to gender, age, educational level, etc. This heterogeneity implies that origin-specific patterns of correlation across all potential locations exist in the stochastic component of utility.

However, the absence of future attractiveness of alternative locations which has an impact on choices of moving between j and k can still bias the estimation (Bertoli and Fernández-Huertas Moraga 2013). More specifically, assume that migration between Suzhou and Shanghai increases because the movers have worse expectations about their job prospects in Suzhou, and these worsening prospects are correlated with their worsening job prospects in Nanjing. Then, if the influence of alternative locations is not considered, the increase in the bilateral migration to Shanghai would be wholly attributed to worsening job prospects in Suzhou, resulting in overestimation. In other words, failure to account for multilateral resistance to migration might entail a violation of the IIA hypothesis underlying the discrete choice model discussed above.

Ideally, the solution to multilateral resistance to migration is [Pesaran's \(2006\)](#) common correlated effects (CCE) estimator if the cross-sectional and longitudinal dimensions of the panel are large enough. Yet our dataset does not meet this computing demand, so we follow the less data-demanding approach used in a wide range of applied migration literature (e.g. [Ortega and Peri 2013](#); [Beine and Parsons 2015](#); [Royuela and Ordóñez 2018](#)).

We change equation (17), first by replacing monadic fixed effects of origin and destination with dyadic fixed effects of origin-destination, as follows:

$$M_{ijk,t} = \alpha + \beta_1 X_{it} + \beta_2 \text{Distance}^T(\text{Prospect}_{ijk,t}) + \beta_3 \text{Distance}(Z_{ijk,t-1}) \\ + \gamma_t + \gamma_{jk} + \gamma_s + \epsilon_{ijk,t} \quad (18)$$

This specification is similar to the above but allows us to control for deterministic effects for each pair of cities. γ_{jk} can capture specific bilateral migration relationships between j and k , such as geographic distance, migration costs, or historic migration networks. Other variables remain the same, as above.

One of the most common approaches to dealing with multilateral resistance to migration is to apply dyadic fixed effects of origin-time, as follows:

$$M_{ijk,t} = \alpha + \beta_1 X_{it} + \beta_2 \text{Distance}^T(\text{Prospect}_{ijk,t}) + \beta_3 \text{Distance}(Z_{ijk,t-1}) \\ + \gamma_{jt} + \gamma_k + \gamma_s + \epsilon_{ijk,t} \quad (19)$$

where γ_{jt} is a vector of origin dummies for each year. All other variables remain the same as in the main specification. In terms of [Royuela and Ordóñez \(2018\)](#); [Ortega and Peri \(2013\)](#), this method enables us to control for all the push determinants of migration decisions and particularly, multilateral resistance derived from heterogeneity in migration preferences that are constant across destination cities and that vary only by year and city of origin. Previous literature where DV was measured through migration flows also used γ_{jt} to account for the denominator $P_{ijj,t}$ ([Beine, Bertoli, and Fernández-Huertas Moraga 2016](#)).

Another common approach is to use dyadic fixed effects of destination-time in the place of origin-time, as follows:

$$M_{ijk,t} = \alpha + \beta_1 X_{it} + \beta_2 \text{Distance}^T(\text{Prospect}_{ijk,t}) + \beta_3 \text{Distance}(Z_{ijk,t-1}) \\ + \gamma_{kt} + \gamma_j + \gamma_s + \epsilon_{ijk,t} \quad (20)$$

where γ_{kt} is a vector of destination dummies for each year. All other variables remain the same as in the main specification. As explained in [Beine and Parsons \(2015\)](#) and [Royuela and Ordóñez \(2018\)](#), this strategy allows us to control for all the pull determinants of migration decisions and dynamic resistance derived from heterogeneity in the future attractiveness that are constant across origin cities and that vary only by year and city of destination.

Based on equations (19) and (20), we replace the monadic fixed effects of industrial sector with dyadic fixed effects of industry-time because our trending indicator rests on

sectors:

$$\begin{aligned}
 M_{ijk,t} &= \alpha + \beta_1 X_{it} + \beta_2 \text{Distance}^T(E_Trending_{ijk,t}) + \beta_3 \text{Distance}(Z_{ijk,t-1}) \\
 &\quad + \gamma_{jt} + \gamma_k + \gamma_{st} + \epsilon_{ijk,t} \\
 M_{ijk,t} &= \alpha + \beta_1 X_{it} + \beta_2 \text{Distance}^T(E_Trending_{ijk,t}) + \beta_3 \text{Distance}(Z_{ijk,t-1}) \\
 &\quad + \gamma_{kt} + \gamma_j + \gamma_{st} + \epsilon_{ijk,t}
 \end{aligned} \tag{21}$$

where γ_{st} is a vector of industry dummies for each year, capturing heterogeneity in migration preferences that vary by year and industrial sector, such as emerging jobs.²² All other variables remain the same as in the main specification.

As we have seen, the less data-demanding solution to multilateral resistance to migration relies on utilising various structures of fixed effects. Nevertheless, a fixed effects estimator can be significantly biased in non-linear models. Beck (2018) demonstrates that the larger the number of fixed effects, the stronger bias imposed on fixed effects logit models. In contrast, the critique of using a linear probability model (LPM) is usually twofold: a) predicted probabilities might be negative or above 1, such as 1.2 or -0.4 , that are unrealistic, and b) the dichotomous DV renders heteroskedasticity which violates one of the OLS assumptions that all disturbances have the same variance. Yet the heteroskedasticity can be addressed with heteroskedasticity-consistent robust standard errors, while the first concern is most often why the LPM is not preferred. Horrace and Oaxaca (2006) shows that the bias and inconsistency of LPM increase with a greater proportion of predicted probabilities falling outside the unit interval. In other words, if all predicted probabilities fall between 0 and 1, the linear probability estimator can be unbiased and consistent.

In our sample, 8.3% (6.2%) of observations present a negative predicted migratory probability drawn from equation (17). We, thus, report estimates of the sub-sample where all predicted probabilities are within the unity as a supplement. Further, we consider a multilevel logistic regression as part of the robustness check. It enables us to perform logistic regression along with controlling for unmeasured context-specific influences on potential migrants that is in accord with the theoretical perspective that people from the same place of origin tend to behave more similarly than do individuals from other places due to a variety of spatial and socio-cultural proximity.

Let $\pi_{it} = \text{Pr}(M_{ijk,t} = 1)$, our two-level logistic random intercept model is formulated as follows and purely an empirical counterpart of equation (10):

$$\begin{aligned}
 \log\left(\frac{\pi_{it}}{1 - \pi_{it}}\right) &= \alpha + \beta_1 X_{it} + \beta_2 \text{Distance}^T(\text{Prospect}_{ijk,t}) + \beta_3 \text{Distance}(Z_{ijk,t-1}) \\
 &\quad + \gamma_t + \mu_C
 \end{aligned} \tag{22}$$

where μ_C is assumed to be i.i.d normally distributed with a zero mean and level-2 variance σ_C , accounting for the effects of being in city group C on the log-odds that $M_{ijk,t} = 1$. β_1 is still level-1 unknown parameters, and β_2 , β_3 and β_4 are level-2 parameters to be estimated. The index C can be either an origin city or a destination city. Later, we estimate both.

We can extend the two-level model to a three-level model if we postulate that within the same city group, the type of contextual information that people access and the way

they are impacted differ across education levels. In this case, city effects μ_C become level-3 and a new level-2 random intercept, i.e. μ_{Cv} , is added to equation (22). The index v denotes the year of schooling.

In addition, Wintoki, Linck, and Netter (2012) show that the generalised method of moments (GMM) estimation yields better results than a fixed effects estimator for at least two potential issues of endogeneity: unobserved heterogeneity and simultaneity. On the basis of Hansen (1982), the two-step system GMM is further developed by Blundell and Bond (1998) where lagged first differences are utilised as instruments in the level equation at the cost of an additional assumption that first differences of instrument variables are uncorrelated with unobserved unit-specific heterogeneity. We, thus, further continue our analysis using a system GMM estimator.²³ The model is almost the same as equation (16) but with an inclusion of industry fixed effects.

3.3. Data description

3.3.1. Data source

The quasi-panel is created by combining a nationally representative cross-sectional micro-data, the 2017 China Household Finance Survey (CHFS), with city-level longitudinal statistics during 1997–2017, retrieved from the China Data Institute. This approach is not uncommon for applied research.²⁴ The CHFS is one of the very few nationwide surveys that allow us to identify where migrants are from and where they settle at the city level and the year in which their migration occurs.²⁵ After comprehensive data cleaning and compiling, our panel eventually contains 10,254 migrants and 56,173 natives. For details of the data preparation, please see Appendix 4.

It is worth noting again that all individual and household variables were retrieved from the 2017 wave. After merging the cross-sectional survey data with the longitudinal statistics, we re-calculated the age and years of schooling of each individual for each year and only retained observations of ages between 16 and 65. As a result, we finally have three time-varying variables at this level, that is: age, years of schooling, and family migration network. As migrants are defined as those who have moved across prefecture boundaries, and to focus on the decision-making problem, our binary DV is only equal to 1 in the year of migration and to 0 at all other points in time. Furthermore, we distinguished individuals who had moved and transferred their Hukou to destination cities from natives and specified them as migrants. As migrants in China are often measured by the separation of the Hukou and resident places, people who transferred the Hukou to their cities of destination are treated as natives in the majority of migration studies (Zhang et al. 2020). However, ignoring those migrants with Hukou transfer leads to an underestimate of the scale of migration.

3.3.2. Descriptive statistics

In our RUM model, we discuss that the expected utility $A_{kt+1}(I)$ is conditional on the industry in which job categories that individual i seek are mainly based. From the CHFS, we learn each individual's industry group, if he or she has a formal job. Table 1 classifies the industry groups based on the Chinese national standard number 'GB/T 4754' and presents corresponding statistics of each group. As shown, at the aggregate level, the primary sector has the lowest proportion of labour, as opposed to the tertiary sector. Among sub-sectors, 21.3% of individuals in our sample were employed in

Table 1. Industrial classification & statistics.

Industry Group	Num. of Obs.	Main Sector
Farming, Forestry, Animal Husbandry	2695	Primary Sector
Mining and Quarrying	488	Secondary Sector
Manufacturing	7765	Secondary Sector
Electric Power Gas and Water Production and Supply	1184	Secondary Sector
Construction	3517	Secondary Sector
Wholesale and Retail Trade	3294	Tertiary Sector
Transportation Storage Post and Telecommunications	2531	Tertiary Sector
Hotel and Catering Services	1959	Tertiary Sector
Information Transmission, Software and Information Technology	870	Tertiary Sector
Banking and Insurance	1015	Tertiary Sector
Real Estate	466	Tertiary Sector
Leasing and Business Services	333	Tertiary Sector
Scientific Research, Technical Service and Geologic Prospecting	234	Tertiary Sector
Management of Water Conservancy, Environment and Public Facilities	455	Tertiary Sector
Social Services	3314	Tertiary Sector
Education	2112	Tertiary Sector
Health, Social Security and Social Welfare	1562	Tertiary Sector
Culture, Sports and Entertainment	613	Tertiary Sector
Public Management and Social Organisation	1942	Tertiary Sector
<i>Total</i>	36,349	

Source: Authors' calculation using CHFS.

Manufacturing, followed by Construction (9.7%), Social Services (9.1%) and Wholesale and Retail Trade (9%).

The trending indicator is eventually quantified by the statistics of the three main sectors. Approximately 25,000 natives and 3000 migrants in our sample do not provide industry groups of their jobs because they are non-employee workers. Thus, total employment statistics are used for this group instead of sectoral statistics. Accordingly, we have four sets of industry-specific dummies applied in our estimation (primary, secondary, tertiary and total). Furthermore, Figure 4 is the geographic distribution of emigrants employed in the secondary sector moving across province boundaries, from which we see five Chinese provinces suffered the greatest labour outflux: Sichuan, Hubei, Anhui, Hunan, and Guangdong. Similarly, Figure 5 is the plot looking at the tertiary sector. Here, Guangdong still sent out the most migrants followed by Sichuan, Hubei, Anhui, Hunan, and Liaoning. In Figure A1, we also graphically present where migrants are from, defined by the prefecture boundary.

Table 2 contains descriptive statistics of the survey data and city statistics, respectively, as well as the DV, the trending indicator ($Distance^T(E_Trending_{ijk,t})$) and the wage indicator in the final quasi-panel.²⁶ More specifically, migrants in our sample were on average 11 years younger, had 1.4 more years of schooling, and obtained 6800 Yuan more in annual wages after-tax than did natives. The gender ratio of females to males among natives was 46:54, but 42:58 among migrants. As to the Hukou type, 44% of natives, as opposed to 34% of migrants, had an urban Hukou.²⁷ Family sizes did not seem to differ between natives and migrants (1.87 versus 1.89), but fewer migrants were in a marital relationship. Self-evaluated health status shows that migrants overall evaluated their health as slightly better than that of natives.

Moreover, compared to natives, migrants were mostly from economically underdeveloped cities where the GDP per capita was on average 4600 Yuan less in 2000 and

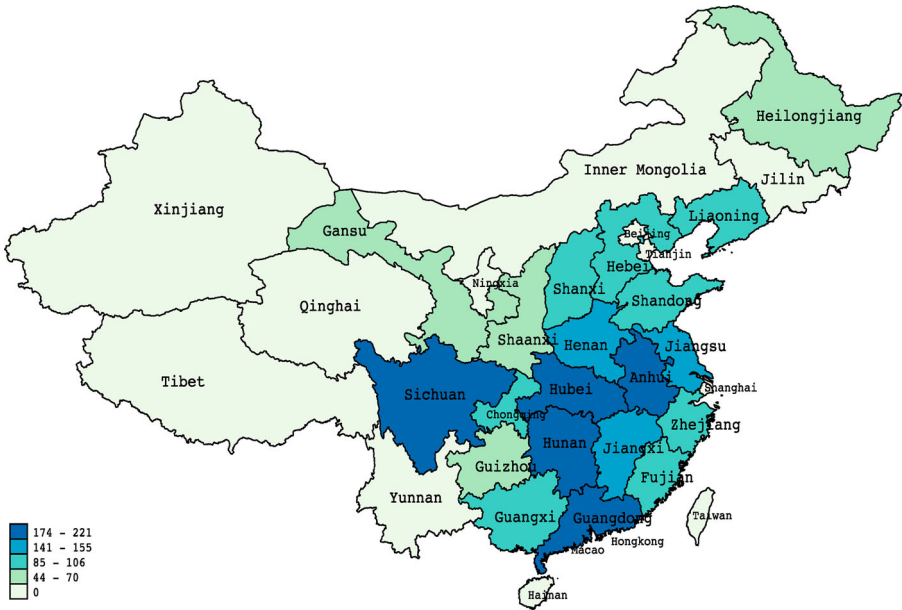


Figure 4. Province boundary map – emigration distribution (secondary sector).
Author's elaboration using CHFS.



Figure 5. Province boundary map – emigration distribution (tertiary sector).
Author's elaboration using CHFS.

12,500 Yuan less in 2010 than natives' resident cities. Generally, migrants moved to economically developed cities where the GDP per capita was 9800 Yuan more in 2000 and 50,000 Yuan more in 2010 than their areas of origin. Likewise, destination cities usually had 9% higher ratios of the tertiary sector to the total than cities of origin and three to

Table 2. Descriptive statistics.

Variable	Description	Obs.	Mean	Std. Dev.	Min	Max
<i>CHFS 2017</i>						
migrant	If native = 0; migrant = 1	75,099	0.17	0.38	0	1
gender	Female = 0; Male = 1	75,099	0.55	0.50	0	1
age	Age in 2017	75,099	47.82	17.66	16	85
education	Educational level: no schooling = 1, PhD = 9	74,894	3.48	1.75	1	9
marriage	Married = 1; Otherwise = 0	72,835	0.77	0.42	0	1
hukou_type	Urban Hukou = 1; Rural = 0	74,961	0.41	0.49	0	1
health_status	Health degree: very good = 1, very bad = 5	75,072	2.52	1.03	1	5
hh_income	Log of household income	75,099	13.89	0.36	−9.21	15.61
family_size	The number of family members living together	12,214	1.88	1.28	1	10
wage	After-tax wage last year	24,736	37,533.76	45,066.41	20	3,100,000
<i>City Statistics</i>						
all_GR	Growth rates in all sectors	5522	0.0208	0.4284	−0.9673	18.36
prima_GR	Growth rates in the primary sector	5230	0.1581	3.55	−1	120.96
second_GR	Growth rates in the secondary sector	5738	0.0132	0.4543	−0.9267	22.72
tertiary_GR	Growth rates in the tertiary sector	5738	0.0117	0.3121	−0.9395	7.78
ppDen	Log of population density	5667	5.85	1.01	1.55	9.55
lngdppc	Log of GDP per capita	5492	9.73	0.9582	7.01	12.58
coop	The number of enterprises above the designated size per 10,000 persons	5425	2.53	3.32	0.12	36.29
medical	Number of beds in hospitals per person	5705	0.0033	0.0016	0	0.0189
highEdu	Number of higher educational institutions per 10,000 persons	5739	0.0155	0.0194	0	0.1197
tertiaryRatio	The ratio of output values of the tertiary sector	5780	0.6529	0.1553	0.2029	0.9897
CHRI	The Hukou registration stringency index	2167	0.6614	0.3370	0.1331	2.63
<i>Merged</i>						
migrate	= 1 at the time of migration; = 0 before and afterwards	1,128,624	0.0084	0.0914	0	1
distance_ETrend	The ratio of employment prospects at destination to origin	969,998	−7.37e−06	0.0045	−0.6976	0.4884
distance_wage	The ratio of wage prospects at destination to origin	1,050,567	0.0387	0.2940	−5.94	7.75
pioneer	If any pioneer migrants within families = 1; No = 0	1,128,624	0.0237	0.1522	0	1

Notes: Approx. 10,000 observations in the 2017 CHFS survey cannot match with city statistics because their resident cities are not included in the statistics. 1,466 households answered with zero or even negative annual income due to debts and losses on investment, among others. The variable 'distance_ETrend' and 'distance_wage' are the transformed, as defined in equations (14) and (15). The statistical currency is Yuan.

Source: Authors' elaboration using CHFS, CHRI and China Data Institute (2021).

four times more enterprises above the designated size per 10,000 persons. These destination cities were also on average 41% more densely populated than migrants' cities of origin but with better public medical and higher education resources per capita. In addition, natives' resident cities applied more stringent Hukou registration policies: the CHRI was 0.17 higher before 2013 and 0.28 higher since 2014 than migrants' cities of origin. However, the Hukou barrier did not indeed prevent migrants from moving to higher indexed cities.²⁸ This is mainly because cities setting higher registration barriers were, in general, more economically developed (Zhang, Wang, and Lu 2019). Lastly and importantly, annual differences in wage levels in destination cities were, in general, 3.6% higher than in origin cities (3353.42 versus 3237.40 Yuan) and the distance increased to 27.6% (4241.59 versus 3322.94 Yuan) at the time that migration occurred. Similarly, although the percentage change in growth rates of industrial employment in origin cities fared slightly better than destination cities, for example, 4.7% at origin vs. 3% at potential destination in 2000 and 1.6% at origin vs. 0.9% at potential destination in 2010, the situation completely reversed at the time that migration occurred: 0.8% at origin vs. 2.2% at destination, on average.

4. Estimation

4.1. Results

Results in terms of equations (16)–(21) are reported in Table 3. The most basic result in Column (1) indicates that a 10% change in the ratio of employment prospects (annual changes in sector-specific employment growth) at destination to origin is positively associated with an increase of 0.876 percentage points in migratory probabilities. It is worth noting that small R^2 values here are endemic to and a direct result of our discrete choice setting. Then, we added individual, household and city covariates to the second model, as seen in Column (2). The effects of employment prospects increase considerably, from 0.876 to 1.933 percentage points, and additionally, having any pioneer migrants within families prior to the move leads to an increase of 6.67 percentage points in migratory probabilities. In addition, we estimate the same models using the distance in employment growth rates, i.e. $GR_{ik,t} - GR_{ij,t}$, and find that it has little impact on migration, as opposed to the trending indicator.²⁹

From Column (3) onward, we begin to eliminate unobserved heterogeneity where various options of fixed effects were added, and in Column (4), we included the CHRI (henceforth, the Hukou index). As reported in Column (3) and Columns (5)–(9), the magnitude of the coefficients of the ratio of sector-based employment prospects in cities of destination to cities of origin is between 1.321 and 1.788 percentage points. Similarly, the magnitude of the coefficients of the family migration network is between 5.96 and 6.66 percentage points. Column (4) can be seen as a subpopulation analysis, because only 120 cities have Hukou indices. Results indicate that, after controlling for the Hukou registration stringency, the influence employment prospects exert on migration climbs to 1.636 percentage points. And a unit increase in the distance in Hukou indices raises the probability of migration by 2.22 percentage points. The stringency gap *per se* would not attract labour, but projects regional disparities in economic development and urbanisation.³⁰ However, the interaction effect between the Hukou index and employment prospects is just statistically significant at the 10% level.

Table 3. Determinants of migration decisions (1997–2017): employment prospects with fixed effects.

<i>No restrictions</i>									
	OLS		Multilateral resistance to migration						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
distance_ETrend	0.0876*** (0.0297)	0.1933*** (0.0290)	0.1408*** (0.0330)	0.1636*** (0.0579)	0.1491*** (0.0329)	0.1721*** (0.0576)	0.1321*** (0.0321)	0.1354*** (0.0314)	0.1788*** (0.0576)
distance_CHRI				0.0222** (0.0087)					
distance_(ETrendXCHRI)				0.2212* (0.1129)					
pioneer		0.0667*** (0.0043)	0.0619*** (0.0035)	0.0576*** (0.0063)	0.0666*** (0.0045)	0.0596*** (0.0033)	0.0614*** (0.0036)	0.0617*** (0.0037)	0.0601*** (0.0034)
Ind. Controls		Y	Y	Y	Y	Y	Y	Y	Y
City Controls		Y	Y	Y	Y	Y	Y	Y	Y
Constant	0.0090*** (0.0009)	0.0622*** (0.0045)	0.0603*** (0.0041)	0.0406*** (0.0038)	0.0396*** (0.0042)	0.0607*** (0.0042)	0.0600*** (0.0042)	0.0601*** (0.0042)	0.0606*** (0.0041)
Time FE	Y	Y	Y	Y	Y				
Industry FE			Y	Y	Y	Y	Y		
Origin FE			Y	Y		Y			Y
Destination FE			Y	Y			Y	Y	
Pairs of cities FE					Y				
Origin-year FE							Y	Y	
Dest-year FE						Y			Y
Industry-year FE								Y	Y
R ²	0.0014	0.0459	0.0534	0.0509	0.0842	0.0721	0.0697	0.0696	0.0721
Obs	969,998	749,219	729,965	408,877	729,960	729,769	729,725	748,981	749,034

Notes: Complete results of columns (1)–(9) are reported in Appendix 1 using the untransformed trending indicator. Standard errors shown in parentheses are clustered at the destination city. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Source: Created by authors using CHFS, CHRI and China Data Institute (2021).

In Table 4, we restricted observations to those whose predicted probabilities are within the unity. Except for Column (2), all regressions involved more than 90% of the observations. Estimates of employment prospects in Columns (1), (2), (4) and (7) are found to be larger than their counterparts reported in Columns (3), (4), (6) and (9) of Table 3. In contrast, Columns (3), (5) and (6) show slightly smaller effects of employment prospects on migration decisions compared to Columns (5), (7) and (8) of Table 3. In sum, the magnitude of the coefficients of the ratio of sector-based employment prospects at destination to origin is between 1.281 and 2.185 percentage points. Furthermore, the interaction term is statistically insignificant. Given these results, we conclude that our linear probability estimator performs acceptably.³¹

Regarding estimates of control variables reported in Table A1, we see that migratory probabilities increase for people who are male, unmarried, younger, or more educated. As we focus on labour migrants, our findings are distinguished from those studies that do not primarily analyse labour migration. For example, in this study, men are found to be more likely to migrate than women, whereas an opposite story is not uncommon among other migration scenarios, such as permanent migration (e.g. Meng 2020; Zhang et al. 2020). As for city covariates, the distance in income is found statistically insignificant in Columns (3)–(9) when we further control for heterogeneity in migration preferences. Instead, two variables are consistently significant, that is, the distance in the provision of healthcare and business density. As expected, abundant medical resources positively drive migration, while the negative role of business density in migration seems surprising.³²

In Table A4 we report results of wage prospects. Columns (1)–(4) show that a 10% change in the ratio of wage prospects at destination to origin is positively correlated with an increase of 0.237, 0.041, 0.039, and 0.033 percentage points in migratory probabilities. However, once we include both origin and destination fixed effects, the wage indicator turns statistically insignificant. This manifests why including bilateral fixed effects are important here because, by using only origin (destination) city fixed effects, certain time-invariant or slowly varying unobserved factors across cities on the other side could still be correlated with the wage indicator. Considering wage prospects are not consistently statistically significant, we do not perform robustness checks for it in the next section.

4.2. Robustness checks

4.2.1. Multilevel logistic regression

Results of multilevel logit models are reported in Columns (1)–(6) of Table 5. In Column (1), we consider cities of origin as the higher level to control for origin-specific factors affecting the probability of migration. As we see, the intra-class correlation coefficient (ICC) is merely 0.0432, indicating that observations within the same place of origin are different from each other, so we switched to Columns (2) and (3) where cities of destination and origin-destination pairs are treated as level-2, respectively.³³ Here, the ICC is 0.3355 for destination and 0.2783 for pairwise cities, presenting substantial evidence of clustering, and their coefficients are smaller than in Column (1). In other words, observations within the same destination (origin-destination pairs) have a much lower degree of variability compared to their origin-nested counterparts.³⁴ As destination-nested models present the largest ICC, we mainly interpret these results.³⁵ Because coefficients in the logistic regression are not as easily interpreted as coefficients in the linear regression, we graphically illustrate marginal effects of the second model with an interval of 0.2 in

Table 4. Determinants of migration decisions (1997–2017): supplement.*Predicted probabilities between 0 and 1*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
distance_ETrend	0.1515*** (0.0454)	0.2562** (0.1291)	0.1438*** (0.0480)	0.1947** (0.0771)	0.1281*** (0.0431)	0.1312*** (0.0428)	0.2185*** (0.0776)
distance_CHRI		0.0292*** (0.0097)					
distance_(ETrendXCHRI)		0.1933 (0.1952)					
pioneer	0.0625*** (0.0035)	0.0564*** (0.0059)	0.0666*** (0.0046)	0.0610*** (0.0033)	0.0626*** (0.0036)	0.0628*** (0.0036)	0.0618*** (0.0033)
Ind. Controls	Y	Y	Y	Y	Y	Y	Y
City Controls	Y	Y	Y	Y	Y	Y	Y
Constant	0.0621*** (0.0042)	0.0431*** (0.0042)	0.0386*** (0.0041)	0.0596*** (0.0043)	0.0614*** (0.0042)	0.0613*** (0.0042)	0.0589*** (0.0042)
Time FE	Y	Y	Y				
Industry FE	Y	Y	Y	Y	Y		
Origin FE	Y	Y		Y			Y
Destination FE	Y	Y			Y	Y	
Pairs of cities FE			Y				
Origin-year FE					Y	Y	
Dest-year FE				Y			Y
industry-year FE						Y	Y
R ²	0.0564	0.0558	0.0878	0.0801	0.0750	0.0748	0.0804
Obs	668,644	306,227	724,519	680,253	678,702	696,327	701,926

Notes: Here, regressions are conditional on the predicted probabilities produced by the models of Columns (3)–(9) reported in Table 3. Standard errors shown in parentheses are clustered at the destination city. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: Created by authors using CHFS, CHRI and China Data Institute (2021).

Table 5. Determinants of migration decisions: multilevel logit and two-step system GMM.

	Two level logit			Three level logit			GMM		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
distance_ETrend	5.3903*** (1.0279)	4.5964*** (0.6027)	3.3363*** (1.0040)	5.3922*** (1.0181)	4.5847*** (0.6120)	3.3363*** (1.0040)	0.3157*** (0.0754)	0.2493*** (0.0569)	0.2751*** (0.0582)
Ind. Control	Y	Y	Y	Y	Y	Y	Y	Y	Y
City Control	Y	Y	Y	Y	Y	Y	Y	Y	Y
Intercept	−1.839*** (0.3101)	−1.5841*** (0.2573)	0.0077 (0.2468)	−1.9504*** (0.3137)	−1.7681*** (0.2857)	0.0077 (0.2468)	0.0610*** (0.0046)	0.0590*** (0.0046)	0.0590*** (0.0046)
Level 2 var.	0.1484 (0.0372)	1.6611 (0.1671)	1.2684 (0.1009)	0.1869 (0.0312)	0.2416 (0.0455)	1.99e−33 (3.38e−34)			
Level 3 var.				0.1054 (0.0398)	1.6046 (0.1685)	1.2685 (0.1009)			
ICC	0.0432 (0.0104)	0.3355 (0.0224)	0.2783 (0.0160)	0.0816 (0.0112)	0.3595 (0.0214)	0.2783 (0.0160)			
Nest	origin	destination	pair	origin	destination	pair			
sub-Nest				education	education	education			
Time FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE							Y	Y	Y
Num. of instruments							182	180	201
AR(2)							0.724	0.790	0.790
Hansen's J test							0.349	0.263	0.404
<i>Difference-in-Hansen tests</i>									
GMM instruments for levels – Excluding group							0.114	0.108	0.154
GMM instruments for levels – Difference (null H = exogenous)							0.451	0.355	0.578
GMM instrument for distance_trend – Excluding group							0.520	0.354	0.526
GMM instrument for distance_trend – Difference (null H = exogenous)							0.114	0.192	0.173
Obs	749,219	749,219	749,219	749,219	749,219	749,219	729,965	729,965	729,965

Notes: In Appendix 1, we report complete results using the untransformed trending indicator. AR(2) is the Arellano-Bond test for second-order serial correlation with the null hypothesis of no serial correlation in disturbances. Hansen's J test is a test of over-identifying restrictions (in other words, the overall validity of the instruments) using J statistic of Hansen (1982). Likewise, the difference-in-Hansen test is designed to test the validity of subsets of the instruments. Endogenous and predetermined variables are instrumented with their corresponding second- and third-order lags in column (5)–(6). Three more orders of lags are added in column (7). The trending indicator, the income (GDP per capita), and the share of the tertiary sector are treated as endogenous in column (5), while other city-level covariates are treated as predetermined (not strictly exogenous). In column (6)–(7), the population density and business density are additionally treated as endogenous. Moreover, individual and household covariates are specified as exogenous. The Windmeijer correction (Windmeijer 2005) is used in the GMM estimation, and corresponding standard errors are clustered at the destination city. Besides, robust standard errors are also applied in multilevel logit models. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: Created by authors using CHFS and China Data Institute (2021).

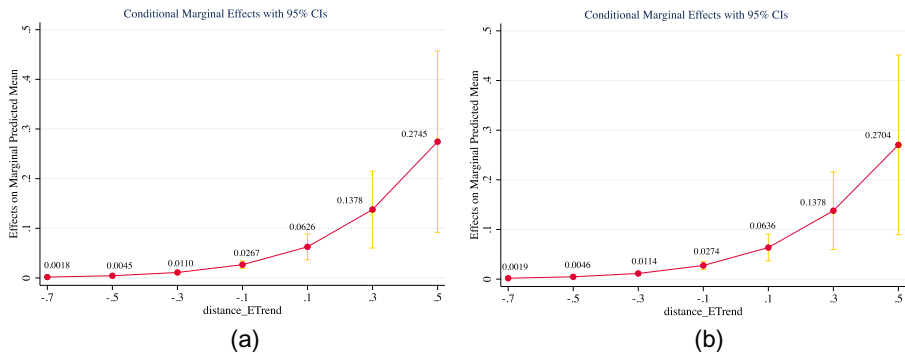


Figure 6. Marginal effects of destination-nested model. (a) Two-level random intercept model and (b) Three-level random intercept model.

Notes: All results are statistically significant at the 1% level.

Created by authors using CHFS and China Data Institute (2021).

Figure 6(a), holding all other variables at mean. The plot clearly shows an increasing trend that, when the ratio of employment prospects at destination to origin gets larger, its effects on motivating migration become increasingly stronger.

It has been widely demonstrated that education levels have a considerable impact on the propensity for migration and location preferences (e.g. Meng 2020; Fu and Gabriel 2012). Thus, we treated years of schooling as a new level, subordinate to cities. In other words, people with identical educational attainments nested in the same locations are supposedly much more similar to each other than their fellow migrants who received a higher or lower level of education. As can be seen, the ICC becomes slightly larger in Columns (4)–(5) than in Columns (1)–(2), whereas almost no change arises from adding the level of education under pairwise nests. It should be noted that cities are level-2 in Columns (1)–(3), but level-3 in Columns (4)–(6). Hence, the variance between cities is 0.1484, 1.6611, and 1.2684 in two-level logit models and 0.1054, 1.6046, and 1.2685 in three-level logit models. Interestingly, the distinction between education groups is even greater than the extent to which that the grouping of origin can account for, as opposed to Column (6), where the variance between level-2 groups within the same origin-destination pairs is nearly zero. The fifth model gives an in-between result: its level-2 variance is around one-seventh of the level-3. We further plot its marginal effects, holding all covariates at mean. As seen in Figure 6(b), the marginal effects are also continuously increasing.

In Table A2, we report the complete regression results of Table 5. The results of covariates vary depending on the definition of nests, except for the distance in income, which is consistently statistically significant across all models. The distance in the provision of healthcare (business density) is found positive (negative) at the 5% (10%) level in pairwise nested models but not in others. In contrast, the distance in population density and the share of the tertiary sector exhibit positive impacts on increasing the probability of migration in destination-nested models.³⁶

4.2.2. Two-step system GMM

In Columns (7)–(9) of Table 5, we report two-step system GMM estimates with different lag and instrument strategies. We adopted two types of fixed effects: time and industry. The former is aimed at absorbing any instant shock exposed to all units, while

the latter is concerned to ensure that variations of employment prospects are estimated within sectors. The first GMM model is the baseline where we utilised two orders of lags to instrument the endogenous and predetermined variables. We initially treated the trending indicator, the income, and the share of the tertiary sector as strictly endogenous, while all other city covariates were treated as predetermined. The coefficient of the trending indicator shows that a 10% change in the ratio of sector-based employment prospects at destination to origin causes an increase of 3.157 percentage points in migratory probabilities. Then, we applied the same lag strategy but, additionally, treated the population density and business density as endogenous variables. As a result, the coefficient is still statistically significant at the 1% level but relatively smaller than the previous result. We kept this instrument strategy but added three more orders of lags to instrument our variables. Adding higher orders of lags can help remove serial correlation but may also impose the weak instrument problem, so we did not include more lags. We learn from Column (9) that the magnitude of the coefficient of our trending indicator is a bit larger but falls exactly between the values of coefficients in Columns (7) and (8). Moreover, the results here are larger than the fixed effects estimates reported in Table 3 and, interestingly, the model of Column (9), where we included the Hukou index and confined observations to having a within-unity predicted probability, produces the closest estimate (2.562 percentage points in Table 4).

As seen in Table A2, half of the city-level control variables are still statistically insignificant in Column (7). While in Columns (8)–(9), the majority appears to be statistically significant, as opposed to the business density, which was initially found to have an impact, but then became negligible when it was treated as endogenous. The second and third GMM models are better than the first in dealing with unobserved heterogeneity, their results are thus more reliable. According to these results, we find that the distance in income, the provision of healthcare, the provision of higher education and population density have positive effects on driving migration. In contrast, the distance in the share of the tertiary sector negatively affects migratory probabilities. The relationship between the distance in population density and migration uncovered here can be attributed to the fact that migrants are attracted to large cities that inevitably are densely populated (Chen and Fan 2016). For the latter, a possible explanation is that considering, in our sample, 39% of total migrants worked in the tertiary sector, cities with a relatively lower share of the tertiary sector are more likely to present a faster pace of growth.³⁷

The results of Hansen's J test for over-identification are well above 0.25, a threshold suggested by Roodman (2009), but far from 1, pointing to a 34.9%, 26.3% and 40.4% chance of a type one error if the null is rejected, and no issue of instrument proliferation. The null hypothesis of the Arellano–Bond test is not rejected, indicating no second-order serial correlation in disturbances.³⁸ By further checking the difference-in-Hansen test for the validity of subsets of instruments, the differenced models are evidenced as dynamically complete, implying the instruments used in the level models are valid. We also report test results of GMM instruments of the trending indicator, from which we can conclude that its corresponding specified instruments used in the level models are exogenous.

5. Discussions and concluding remarks

In this paper, we investigated the effects of job prospects on individual migration decisions across Chinese prefecture boundaries. To this end, we assembled a unique

quasi-panel based on the 2017 China Household Finance Survey and the prefecture city statistics between 1997 and 2017. By accounting for city-level bilateral variations in parallel with individual and household characteristics, we filled gaps in the existing Chinese migration literature in two aspects: (a) migration decisions at the city level are quite scarce due to data limitations, and among them, individual and household characteristics are always absent in analysing regional longitudinal effects, and (b) the previous models that have controlled for these characteristics rely on micro-data whereby regional factors are either missing, monadic, or at the province level. Equally importantly, Chinese migration research is desperately lacking in visionary migration scenarios. By constructing proxy variables for wage and employment prospects, respectively, we enriched economic incentives of labour migration from a forward-looking angle. Further, we theoretically added to the literature of migration decision-making by synthesising the virtues of dynamic discrete choice modelling framework and the reference-dependence derived from prospect theory. Following it, we drew corresponding empirical specifications and applied various monadic and dyadic fixed effects to address multilateral resistance to migration. Further, we considered multilevel logistic regression and two-step system GMM estimation for the robustness check. In sum, this study, acclimatised to the new migration pattern that provincial-level migration has been less popular, deepened the understanding of relationships between regional employment structures and labour mobility in both level and scope.

Wage prospects are influential to migration decisions when we control for unilateral fixed effects, while results become statistically insignificant once bilateral fixed effects are involved. Thus, our primary findings are that a 10% increase in the ratio of employment prospects in cities of destination to cities of origin raises the probability of migration by 1.281–2.185 percentage points, and the effects tend to be stronger when the scale of the ratio is larger. Having a family migration network causes an increase of approximately 6 percentage points in migratory probabilities. Additionally, labour migrants are more likely to be male, unmarried, younger, or more educated. Here, the message our findings deliver is simple – for small- and medium-sized cities that are less attractive to labour migrants than large cities, concentrating resources to stimulate the acceleration of the employment in certain sectors of interest could help, and downward migration may even happen in this case. Future research could pay attention to fieldwork to learn more personal, subjective beliefs from labour migrants to enrich the evaluative dimension of job prospects.

China's extensive, rapid development of infrastructure and transportation boosts the circulation of human capital and labour forces between regions. A high-speed rail now takes only five hours to complete 1318 kilometres between Shanghai and Beijing. These factors underlie the persistence of high labour mobility into the future. At the same time, migrants were found to be increasingly older and more educated from 2000 to 2015, according to China's Migrant Population Development Reports. Significantly, only 2% of migrants had at least a college degree in 1990, while 25 years later, the percentage is 23.3%. Skilled migrants tend to prioritise career prospects over quality of life in the migration decision-making process (Liu and Shen 2014). This trend well signifies why we should re-examine economic incentives from the angle of job prospects. It is evident that China's current and future cohorts, who are more educated, are more likely to make visionary migration decisions than their elder generations. The likelihood of getting a job and how good it is are the most common questions with which (potential) workers are concerned. A variety of dimensions can contribute to assessing job quality, and what we

have attempted to highlight in this paper is but one. Individual fulfilment and aspirations that are much more personal and sophisticated, as Becker and Teney (2020) discuss, can also affect how people regard their job prospects across regions. Our point hereby is basic and general.

Notes

1. Jobless growth means that an economy is growing at a reasonable rate without the proportionate creation of new jobs.
2. Tasks that rely on well-defined procedures and activities are classified as routine while tasks that require creativity, problem-solving or human interaction are classified as non-routine. Both tasks further subdivide into cognitive or manual skill types.
3. Reference-dependence is a central principle in prospect theory. It holds that people evaluate information gradually as losses and gains relative to certain reference points or a status quo, rather than as a final state of an absolute outcome.
4. The grant of Hukou is associated with the access to a variety of social programmes provided by the regional government, such as the entitlement to undertake the college entrance examination (Gaokao), to social security and even to house purchase. Prefecture-level cities are usually the main administrative unit in China designated to manage the household registration (Hukou).
5. In China, prefecture cities rank below provinces, the highest non-national level administrative unit, and above counties.
6. Migration decisions involving future outcomes across location choices are defined as non-myopic migration (equivalent to 'visionary migration' in this study) in a substantial body of migration literature (e.g. Baldwin 2001; Bertoli, Brücker, and Fernández-Huertas Moraga 2016). In contrast, 'Myopia' is a term referring to migration decisions that only depend on the past and/or current situation.
7. The term 'multilateral resistance to migration' is used to explain influences exerted by the attractiveness of alternative locations on migration rates between any pair of regions.
8. It is also worth noting that the model we developed can be applied to other spatial dimensions, not limited to internal migration or China.
9. Cross-sectional and/or provincial analyses predominate Chinese literature on migration (Su, Tesfazion, and Zhao 2018).
10. The location information is usually provided at the province level, meaning that the sending and receiving cities cannot be identified. Further, although investigators provided city information, most often only cities of destination are known to applied researchers, resulting in that the majority of migration studies focus on analysing the receiving context. Until recently, we have still known less about the reasons that migrants leave their hometowns than why they move to their areas of destination.
11. Prior to being popular among migration studies, gravity models are most often used in research on international trade (e.g. Drenski, Hallren, and Lee 2019; Pettersson and Johansson 2013).
12. In other words, conventionally, the deterministic component is measured through instantaneous, absolute outcomes.
13. Housing tenure, the duration at the current address, and neighbourhood socioeconomic status are the variables used to construct endowment effects.
14. In this paper, we are only interested in the reference dependence, so for the description of other features, such as $N(\cdot)$ is concave (risk-averse) for expected gains when $N''(\cdot) \leq 0$ for $x > 0$ and convex (risk-friendly) for expected losses when $N''(\cdot) > 0$ for $x < 0$, please see Czaika (2015).
15. For the reference-independent version, please see Bertoli, Brücker, and Fernández-Huertas Moraga (2016).
16. This assumption causes very little loss of generality. Please see McFadden (1974) for a detailed discussion.
17. We can elaborate 'deviations from two reference points' with two questions. First, are future situations at origin better than the present? Negative changes triggering loss aversion could strongly encourage emigration. Second, are changes at destination better than at origin? A larger scale of expansion at destination is required particularly when positive changes are also observed in origin cities.

18. Because the trending indicator contains negative values, we applied the so-called ‘started logarithm’ (Tukey 1977), i.e. $\ln(y + c)$ where $c > 0$ is set such that $y + c > 0$ for all y , to transform the variable. We estimated both versions and report results of the untransformed distance predictor in Appendix 1.
19. It should be noted that this indicator measures general wage levels of a city, which does not vary by industry, because not all cities report statistics on sector-based wages and the measurement of such data among the rest is not consistent across cities throughout the sample period.
20. The term ‘enterprises above the designated size’ was first used in China in 1996 and defined by the National Bureau of Statistics. Before 2011, firms with an annual output of 5 million Yuan or more were counted as enterprises above the designated size. The benchmark has raised up to 20 million Yuan since 2011.
21. The CHRI divided the local Hukou registration policies into four overarching aspects – talent recruitment, general employment, investment and taxation as well as home purchase and then constructed the index for two stages (2000–2013 and 2014–2017).
22. Since the wage indicator is not sector-specific, we do not estimate this group of equations for it.
23. It is unreasonable to assume that the previous migration decision affects the present migration decision. We therefore do not consider adopting a lagged DV.
24. To give an example, Schmidt-Catran and Spies (2016) exploited hybrid methods to investigate cross-sectional as well as longitudinal effects of migration on native Germans’ support for welfare.
25. Another option is the CMDS, however, we do not have permission to access it.
26. We acknowledge several limitations here. Firstly, as a quasi-panel, most of the individual and household variables are time-invariant. Although the CHFS has four waves, about one-third of migrant samples were new to 2017, and earlier waves still do not have information prior to 2011. Secondly, the migration history we can learn is incomplete. It is possible for migrants to move several times within a year, however, we can only identify their most recent destination. Thirdly, industries where which individuals seek jobs depend on their most recent employment. Mismatch between what we learn from the 2017 wave and actual examples that individuals evaluated is possible. Otherwise, examining job prospects at the sub-industry level would be even more interesting.
27. Prefecture cities in China usually consist of several districts and counties. They have the dual functions of administering both rural and urban areas. An increasing population with the rural Hukou lives in urban areas because of local urbanisation and within-prefecture migration (Song 2014).
28. Actually, their cities of destination were on average 0.37 higher before 2013 and 0.49 higher since 2014 than their cities of origin.
29. We show the major results in Table A3.
30. The top four indexed cities are Beijing, Shanghai, Guangzhou, and Shenzhen, colloquially known as ‘Bei-Shang-Guang-Shen’, representing the most developed areas in China.
31. After computing the predicted probabilities of each model, we counted the proportion of migrant observations with negative fitted values. The proportions are 1.1%, 0.9%, 0.7%, 2.4%, 1.4%, 1.4%, and 2.4% of the total.
32. With a simple regression, the sign remains negative.
33. ICC is calculated as the ratio of the between-group variance relative to the total variance in the sample. It describes the extent to which observations within city groups are similar to each other.
34. This is the reason, in addition to the CHFS’s sampling design, for us to cluster standard errors at the destination city for all non-hierarchical models (see Colin Cameron and Miller 2015).
35. The pairwise nested model produces results most similar to the fixed-effects models but has a lower ICC than the destination-nested model.
36. The relationship between the distance in the share of the tertiary sector and migration decisions is negative when we account for origin- and destination-specific effects simultaneously as in pairwise nested models, despite being statistically insignificant.
37. Based on the variance inflation factor, the colinearity between the trending indicator and the share of the tertiary sector is very weak (their VIFs are 1.01 and 3.01, respectively). Also, the Pearson’s correlation coefficient is just 0.0021.
38. Despite not being reported, no serial correlation was found in AR(3)–(4).
39. Some households were surveyed previously but lost in the later waves.
40. This question asked ‘in which year did he/she leave the city where his/her Hukou is registered’. All the variables discussed in this report are summarised in Table A5.

41. Variables marked with ‘short’ indicate their inputs are based on the first four digits of the NBS codes. This is to ensure that migrants are identified at the prefecture level.
42. In contrast to Figure A4, we showcase that migrants’ origin, Hukou registration and surveyed cities are all non-identical in Figure A5.
43. Recall that A2001 describes family relationships between respondents and interviewed persons.
44. We discussed how we managed it in detail in the next section.
45. In large cities, obtaining a local Hukou is quite hard for immigrants especially since 2014, so it may take a much longer period for people to successfully transfer their Hukous. However, the number of such observations after cleaning is just 745 and among them, migrants who transferred their Hukous to Tier-1 cities are very few. We also confirm that the statistical significance of our trending indicator holds in the case of excluding those observations.

Acknowledgments

Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The authors would like to thank Vicente Royuela, the anonymous referee, the editor, and session participants at the Chinese Economists Society Annual Conference 2021, the 9th International Conference on Applied Research in Economics and Finance, the 13th Chinese Economic Association and 32th CEA UK Annual Conference, the 46th Symposium of the Spanish Economic Association, the 15th RGS Doctoral Conference in Economics, and the 1st International Workshop on the Chinese Development Model for helpful discussions and comments on the research. Any shortcomings are our own.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie [grant agreement number 838534] and the Horizon Europe [grant agreement number 101086139].

ORCID

Huaxin Wang-Lu  <http://orcid.org/0000-0001-5942-9475>

Octasiano Miguel Valerio Mendoza  <http://orcid.org/0000-0001-9127-8777>

References

- Baldwin, R. E. 2001. “Core-Periphery Model with Forward-Looking Expectations.” *Regional Science and Urban Economics* 31 (1): 21–49. doi:10.1016/S0166-0462(00)00068-5
- Baumann, R., J. Svec, and F. Sanzari. 2015. “The Relationship Between Net Migration and Unemployment: The Role of Expectations.” *Eastern Economic Journal* 41 (3): 443–458. doi:10.1057/eej.2014.44
- Beck, N. 2018, September. “Estimating Grouped Data Models with a Binary Dependent Variable and Fixed Effects: What are the Issues.” arXiv preprint arXiv:1809.06505.
- Becker, R., and C. Teney. 2020. “Understanding High-Skilled Intra-European Migration Patterns: The Case of European Physicians in Germany.” *Journal of Ethnic and Migration Studies* 46 (9): 1737–1755. doi:10.1080/1369183X.2018.1561249
- Beine, M., S. Bertoli, and J. Fernández-Huertas Moraga. 2016. “A Practitioners’ Guide to Gravity Models of International Migration.” *The World Economy* 39 (4): 496–512. doi:10.1111/twec.2016.39.issue-4
- Beine, M., P. Bourgeon, and J. Bricongne. 2019. “Aggregate Fluctuations and International Migration.” *The Scandinavian Journal of Economics* 121 (1): 117–152. doi:10.1111/sjoe.2019.121.issue-1

- Beine, M., and C. Parsons. 2015. "Climatic Factors as Determinants of International Migration." *The Scandinavian Journal of Economics* 117 (2): 723–767. doi:10.1111/sjoe.12098
- Bertoli, S., H. Brücker, and J. Fernández-Huertas Moraga. 2016. "The European Crisis and Migration to Germany." *Regional Science and Urban Economics* 60: 61–72. doi:10.1016/j.regsciurbeco.2016.06.012
- Bertoli, S., and J. Fernández-Huertas Moraga. 2013. "Multilateral Resistance to Migration." *Journal of Development Economics* 102: 79–100. doi:10.1016/j.jdeveco.2012.12.001
- Blundell, R., and S. Bond. 1998. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87 (1): 115–143. doi:10.1016/S0304-4076(98)00009-8
- Chen, C., and C. C. Fan. 2016. "China's Hukou Puzzle: Why Don't Rural Migrants Want Urban Hukou?" *China Review* 16 (3): 9–39. <https://muse.jhu.edu/article/634713>.
- Chung, S. H., M. Zhang, and M. D. Partridge. 2020. "Positive Feedback in Skill Aggregation Across Chinese Cities." *Regional Studies* 54 (11): 1596–1611. doi:10.1080/00343404.2020.1743823
- Clark, W. A. V., and W. Lisowski. 2017. "Prospect Theory and the Decision to Move or Stay." *Proceedings of the National Academy of Sciences* 114 (36): E7432–E7440. doi:10.1073/pnas.1708505114
- Colin Cameron, A., and D. L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372. doi:10.3368/jhr.50.2.317
- Czaika, M. 2015. "Migration and Economic Prospects." *Journal of Ethnic and Migration Studies* 41 (1): 58–82. doi:10.1080/1369183X.2014.924848
- Czaika, M., and C. R. Parsons. 2017. "The Gravity of High-Skilled Migration Policies." *Demography* 54 (2): 603–630. doi:10.1007/s13524-017-0559-1
- De Jong, G. F. 2000. "Expectations, Gender, and Norms in Migration Decision-making." *Population Studies* 54 (3): 307–319. doi:10.1080/713779089
- Drenski, A., R. Hallren, and J. N. Lee. 2019. "Tariff Evasion in Global Trade Data." *The Journal of International Trade & Economic Development* 28 (7): 829–842. doi:10.1080/09638199.2019.1602667
- Foltz, J., Y. Guo, and Y. Yao. 2020. "Lineage Networks, Urban Migration and Income Inequality: Evidence From Rural China." *Journal of Comparative Economics* 48 (2): 465–482. doi:10.1016/j.jce.2020.03.003
- Frey, C. B., and M. A. Osborne. 2017. "The Future of Employment: How Susceptible are Jobs to Computerisation?" *Technological Forecasting and Social Change* 114: 254–280. doi:10.1016/j.techfore.2016.08.019
- Fu, Y., and S. A. Gabriel. 2012. "Labor Migration, Human Capital Agglomeration and Regional Development in China." *Regional Science and Urban Economics* 42 (3): 473–484. doi:10.1016/j.regsciurbeco.2011.08.006
- Ge, P., W. Sun, and Z. Zhao. 2021. "Employment Structure in China From 1990 to 2015." *Journal of Economic Behavior & Organization* 185: 168–190. doi:10.1016/j.jebo.2021.02.022
- Hansen, L. P. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (4): 1029. doi:10.2307/1912775
- Hatzigeorgiou, A., and M. Lodefalk. 2021. "A Literature Review of the Nexus Between Migration and Internationalization." *The Journal of International Trade & Economic Development* 30 (3): 319–340. doi:10.1080/09638199.2021.1878257
- He, C., and P. Gober. 2003. "Gendering Interprovincial Migration in China." *International Migration Review* 37 (4): 1220–1251. doi:10.1111/j.1747-7379.2003.tb00176.x
- Horrace, W. C., and R. L. Oaxaca. 2006. "Results on the Bias and Inconsistency of Ordinary Least Squares for the Linear Probability Model." *Economics Letters* 90 (3): 321–327. doi:10.1016/j.econlet.2005.08.024
- Kahneman, D., and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica*, 47, 263–291. *Econometrica* 47 (2): 263–292. doi:10.2307/1914185
- Kennan, J., and J. R. Walker. 2011. "The Effect of Expected Income on Individual Migration Decisions." *Econometrica* 79 (1): 211–251. doi:10.3982/ECTA4657
- Kőszegi, B., and M. Rabin. 2009. "Reference-Dependent Consumption Plans." *American Economic Review* 99 (3): 909–936. doi:10.1257/aer.99.3.909
- Li, X., E. C.-m. Hui, W. Lang, S. Zheng, and X. Qin. 2020. "Transition from Factor-Driven to Innovation-Driven Urbanization in China: A Study of Manufacturing Industry Automation in Dongguan City." *China Economic Review* 59: Article ID 101382. doi:10.1016/j.chieco.2019.101382
- Li, X., and L. Luo. 2021. "Migration Patterns and Migrant Workers' Consumption." *China Agricultural Economic Review* 13 (4): 781–798. doi:10.1108/CAER-08-2020-0197
- Liang, Z. 2001. "The Age of Migration in China." *Population and Development Review* 27 (3): 499–524. doi:10.1111/padr.2001.27.issue-3

- Liu, Y., and J. Shen. 2014. "Spatial Patterns and Determinants of Skilled Internal Migration in China, 2000–2005." *Papers in Regional Science* 93 (4): 749–771. doi:10.1111/pirs.12014
- Liu, Z., and L. Yu. 2020. "Stay or Leave? The Role of Air Pollution in Urban Migration Choices." *Ecological Economics* 177 (June 2019): 106780–106780. doi:10.1016/j.ecolecon.2020.106780.
- Mayda, A. M. 2010. "International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows." *Journal of Population Economics* 23 (4): 1249–1274. doi:10.1007/s00148-009-0251-x
- McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behaviour." In *Frontiers in Econometrics*, edited by P. Zarembka, 105–142. New York: Academic Press.
- Meng, L. 2020. "Permanent Migration Desire of Chinese Rural Residents: Evidence from Field Surveys, 2006–2015." *China Economic Review* 61 (Jan): 1–15. doi:10.1016/j.chieco.2019.01.001.
- Meng, X., and S. Xue. 2020. "Social Networks and Mental Health Outcomes: Chinese Rural–Urban Migrant Experience." *Journal of Population Economics* 33 (1): 155–195. doi:10.1007/s00148-019-00748-3
- Mu, X., A. G.-O. Yeh, X. Zhang, J. Wang, and J. Lin. 2021. "Moving Down the Urban Hierarchy: Turning Point of China's Internal Migration Caused by Age Structure and Hukou System." *Urban Studies* 59 (7): 1389–1405. doi:10.1177/0042098021100779.
- Ortega, F., and G. Peri. 2013. "The Effect of Income and Immigration Policies on International Migration." *Migration Studies* 1 (1): 47–74. doi:10.1093/migration/mns004
- Peng, L., and S. Swider. 2017. "Migration and Regional Inequality: Changing Characteristics of China's Economic Inequality." *Eurasian Geography and Economics* 58 (1): 89–113. doi:10.1080/15387216.2017.1292144
- Pesaran, M. H. 2006. "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica* 74 (4): 967–1012. doi:10.1111/ecta.2006.74.issue-4
- Petterson, J., and L. Johansson. 2013. "Aid, Aid for Trade, and Bilateral Trade: An Empirical Study." *The Journal of International Trade & Economic Development* 22 (6): 866–894. doi:10.1080/09638199.2011.613998
- Poncet, S. 2006. "Provincial Migration Dynamics in China: Borders, Costs and Economic Motivations." *Regional Science and Urban Economics* 36 (3): 385–398. doi:10.1016/j.regsciurbeco.2005.11.003
- Qu, Y., F. Cai, and X. Zhang. 2013. "Has the 'Flying Geese' Occurred in China? An Analysis of China's Manufacturing Industries from 1998 to 2009." *China Economic Quarterly* 12 (3): 757–776. <https://library.oapen.org/bitstream/handle/20.500.12657/33583/459850.pdf?>
- Ren, G., X. Zhu, N. Heerink, and S. Feng. 2020. "Rural Household Migration in China – the Roles of Actual and Perceived Tenure Security." *China Economic Review* 63 (6): Article ID 101534. doi:10.1016/j.chieco.2020.101534
- Roodman, D. 2009. "How to Do Xtabond2: An Introduction to Difference and System GMM in Stata." *Stata Journal* 9 (1): 86–136. doi:10.1177/1536867X0900900106
- Royuela, V., and J. Ordóñez. 2018. "Internal Migration in a Developing Country: A Panel Data Analysis of Ecuador (1982–2010)." *Papers in Regional Science* 97 (2): 345–367. doi:10.1111/pirs.v97.2
- Şahin, A., A. Tansel, and M. H. Berument. 2015. "Output-Employment Relationship Across Sectors: A Long- Versus Short-Run Perspective." *Bulletin of Economic Research* 67 (3): 265–288. doi:10.1111/boer.2015.67.issue-3
- Schmidt-Catran, A. W., and D. C. Spies. 2016. "Immigration and Welfare Support in Germany." *American Sociological Review* 81 (2): 242–261. doi:10.1177/0003122416633140
- Shrestha, M. 2020. "Get Rich or Die Tryin': Perceived Earnings, Perceived Mortality Rates, and Migration Decisions of Potential Work Migrants from Nepal." *The World Bank Economic Review* 34 (1): 1–27. doi:10.1093/wber/lhz023
- Small, K. A., and H. S. Rosen. 1981. "Applied Welfare Economics with Discrete Choice Models." *Econometrica* 49 (1): 105. doi:10.2307/1911129
- Song, Y. 2014. "What Should Economists Know about the Current Chinese Hukou System?" *China Economic Review* 29: 200–212. doi:10.1016/j.chieco.2014.04.012
- Su, Y., P. Tesfazion, and Z. Zhao. 2018. "Where are the Migrants from? Inter- vs. Intra-Provincial Rural–Urban Migration in China." *China Economic Review* 47: 142–155. doi:10.1016/j.chieco.2017.09.004
- Tan, Y., H. Xu, and X. Zhang. 2016. "Sustainable Urbanization in China: A Comprehensive Literature Review." *Cities (London, England)* 55: 82–93. doi:10.1016/j.cities.2016.04.002.
- Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

- Wang, L., Y. Xue, M. Chang, and C. Xie. 2020. "Macroeconomic Determinants of High-Tech Migration in China: The Case of Yangtze River Delta Urban Agglomeration." *Cities (London, England)* 107 (Sep. 2019). doi:[10.1016/j.cities.2020.102888](https://doi.org/10.1016/j.cities.2020.102888).
- Windmeijer, F. 2005. "A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators." *Journal of Econometrics* 126 (1): 25–51. doi:[10.1016/j.jeconom.2004.02.005](https://doi.org/10.1016/j.jeconom.2004.02.005)
- Wintoki, M. B., J. S. Linck, and J. M. Netter. 2012. "Endogeneity and the Dynamics of Internal Corporate Governance." *Journal of Financial Economics* 105 (3): 581–606. doi:[10.1016/j.jfineco.2012.03.005](https://doi.org/10.1016/j.jfineco.2012.03.005)
- Zhang, J., J. Huang, J. Wang, and L. Guo. 2020. "Return Migration and Hukou Registration Constraints in Chinese Cities." *China Economic Review* 63: Article ID 101498. doi:[10.1016/j.chieco.2020.101498](https://doi.org/10.1016/j.chieco.2020.101498)
- Zhang, J., R. Wang, and C. Lu. 2019. "A Quantitative Analysis of Hukou Reform in Chinese Cities: 2000–2016." *Growth and Change* 50 (1): 201–221. doi:[10.1111/grow.2019.50.issue-1](https://doi.org/10.1111/grow.2019.50.issue-1)
- Zhang, J., and Z. Zhao. 2013. "Measuring the Income-Distance Tradeoff for Rural-Urban Migrants in China." IZA Discussion Paper (7160).
- Zhao, X. 2020. "Migrants and Urban Wage: Evidence from China's Internal Migration." *China Economic Review* 61 (Mar): Article ID 101287. doi:[10.1016/j.chieco.2019.03.006](https://doi.org/10.1016/j.chieco.2019.03.006)
- Zhou, G., G. Chu, L. Li, and L. Meng. 2020. "The Effect of Artificial Intelligence on China's Labor Market." *China Economic Journal* 13 (1): 24–41. doi:[10.1080/17538963.2019.1681201](https://doi.org/10.1080/17538963.2019.1681201)

Appendices

Appendix 1 Regression tables of complete results

Table A1. Determinants of migration decisions (1997–2017): employment prospects with fixed effects

	OLS		Multilateral resistance to migration						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
distance_ETrend	0.0011*** (0.0003)	0.0020*** (0.0003)	0.0014*** (0.0003)	0.0018*** (0.0007)	0.0015*** (0.0003)	0.0017*** (0.0005)	0.0013*** (0.0003)	0.0013*** (0.0003)	0.0018*** (0.0005)
gender		0.0009*** (0.0001)	0.0006*** (0.0001)	0.0005*** (0.0002)	0.0002 (0.0001)	0.0006*** (0.0001)	0.0006*** (0.0001)	0.0006*** (0.0001)	0.0006*** (0.0001)
marriage		−0.0080*** (0.0006)	−0.0074*** (0.0005)	−0.0048*** (0.0006)	−0.0041*** (0.0005)	−0.0072*** (0.0005)	−0.0074*** (0.0005)	−0.0074*** (0.0005)	−0.0073*** (0.0005)
hukou_type		−0.0010** (0.0004)	−0.0004 (0.0003)	−0.0008** (0.0003)	0.0000 (0.0002)	−0.0004 (0.0003)	−0.0004 (0.0003)	−0.0004 (0.0003)	−0.0004 (0.0003)
health_status		−0.0002** (0.0001)	−0.0001 (0.0001)	0.0000 (0.0001)	−0.0001* (0.0001)	−0.0001* (0.0001)	−0.0001 (0.0001)	−0.0001* (0.0001)	−0.0001* (0.0001)
hh_income		0.0000 (0.0001)	−0.0000 (0.0001)	0.0001 (0.0001)	−0.0001** (0.0001)	0.0000 (0.0001)	−0.0000 (0.0001)	−0.0000 (0.0001)	0.0000 (0.0001)
age		−0.0025*** (0.0002)	−0.0023*** (0.0002)	−0.0016*** (0.0002)	−0.0014*** (0.0002)	−0.0023*** (0.0002)	−0.0023*** (0.0002)	−0.0023*** (0.0002)	−0.0023*** (0.0002)
age ²		0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)
schooling		0.0002*** (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)	0.0000 (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)
pioneer		0.0667*** (0.0043)	0.0619*** (0.0035)	0.0575*** (0.0063)	0.0666*** (0.0045)	0.0596*** (0.0033)	0.0614*** (0.0036)	0.0617*** (0.0037)	0.0601*** (0.0034)
distance lngdppc		0.0144*** (0.0023)	0.0003 (0.0044)	0.0034 (0.0062)	0.0038 (0.0045)	−0.0019 (0.0055)	−0.0016 (0.0045)	−0.0020 (0.0045)	−0.0027 (0.0054)
distance_coop		−0.0004 (0.0003)	−0.0007** (0.0003)	−0.0003 (0.0003)	−0.0007** (0.0003)	−0.0008** (0.0004)	−0.0008*** (0.0003)	−0.0008*** (0.0003)	−0.0008* (0.0004)

distance_medical		2.2892**	3.2870***	2.4882**	3.0634***	4.1929***	3.2512***	3.2223***	4.1993***
		(1.0015)	(0.9410)	(0.9847)	(0.8152)	(1.2921)	(1.0355)	(1.0390)	(1.2887)
distance_highEdu		0.1441*	0.0961	−0.0415	0.0404	0.1681*	0.1215	0.1262	0.1716*
		(0.0812)	(0.0922)	(0.1496)	(0.0871)	(0.0941)	(0.0948)	(0.0903)	(0.0886)
distance_ppDen		0.0026	−0.0016	−0.0008	0.0008	−0.0097	−0.0007	−0.0006	−0.0096
		(0.0019)	(0.0084)	(0.0055)	(0.0074)	(0.0075)	(0.0080)	(0.0080)	(0.0074)
distance_tertiaryRatio		0.0024	−0.0348	−0.0259	−0.0380	−0.0306	−0.0335	−0.0368	−0.0320
		(0.0188)	(0.0254)	(0.0342)	(0.0249)	(0.0273)	(0.0253)	(0.0252)	(0.0271)
distance_CHRI				0.0221**					
				(0.0088)					
distance_ETrendXdistance_CHRI				0.0022**					
				(0.0010)					
Constant	0.0089***	0.0622***	0.0604***	0.0405***	0.0396***	0.0607***	0.0600***	0.0601***	0.0606***
	(0.0009)	(0.0045)	(0.0041)	(0.0038)	(0.0042)	(0.0042)	(0.0042)	(0.0042)	(0.0041)
Time FE	Y	Y	Y	Y	Y				
Industry FE			Y	Y	Y	Y	Y		
Origin FE			Y	Y		Y			Y
Destination FE			Y	Y			Y	Y	
Pairs of cities FE					Y				
Origin-year FE							Y	Y	
Dest-year FE						Y			Y
Industry-year FE							Y	Y	Y
R ²	0.0014	0.0459	0.0534	0.0509	0.0842	0.0721	0.0697	0.0696	0.0721
Obs	969,998	749,219	729,965	404,487	729,960	729,769	729,725	748,981	749,034

Notes: Trending indicators estimated here are untransformed, i.e. $E_Trending_{ik,t} - E_Trending_{ij,t}$. Standard errors shown in parentheses are clustered at the destination city. Coefficients are displayed as 0.0000 because the values are smaller than 0.0001. The small R^2 is endemic to and a direct result of our discrete choice setting. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: Created by authors using CHFS, CHRI and China Data Institute (2021).

**Table A2.** Determinants of migration decisions: multilevel logit and two-step system GMM.

	Two level logit			Three level logit			GMM		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
distance_ETrend	0.0533*** (0.0106)	0.0446*** (0.0058)	0.0325*** (0.0104)	0.0534*** (0.0106)	0.0445*** (0.0058)	0.0325*** (0.0104)	0.3157*** (0.0754)	0.2493*** (0.0569)	0.2751*** (0.0582)
gender	0.1153*** (0.0238)	0.0808*** (0.0192)	0.0303* (0.0159)	0.1265*** (0.0243)	0.0950*** (0.0194)	0.0303* (0.0159)	0.0010*** (0.0002)	0.0010*** (0.0002)	0.0010*** (0.0002)
marriage	−0.4273*** (0.0302)	−0.3765*** (0.0346)	−0.4047*** (0.0242)	−0.4110*** (0.0300)	−0.3520*** (0.0340)	−0.4047*** (0.0242)	−0.0078*** (0.0006)	−0.0076*** (0.0006)	−0.0076*** (0.0006)
hukou_type	−0.0421 (0.0350)	−0.0344 (0.0442)	0.0332 (0.0240)	−0.0589* (0.0355)	−0.0567 (0.0447)	0.0332 (0.0240)	−0.0008* (0.0004)	−0.0007* (0.0004)	−0.0008* (0.0004)
health_status	−0.0163 (0.0147)	−0.0083 (0.0135)	0.0004 (0.0112)	−0.0165 (0.0148)	−0.0056 (0.0136)	0.0004 (0.0112)	−0.0002** (0.0001)	−0.0002* (0.0001)	−0.0002* (0.0001)
hh_income	0.0012 (0.0099)	−0.0066 (0.0080)	−0.0148** (0.0075)	−0.0020 (0.0099)	−0.0076 (0.0083)	−0.0148** (0.0075)	0.0001 (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)
age	−0.1944*** (0.0077)	−0.1885*** (0.0080)	−0.1943*** (0.0065)	−0.1930*** (0.0074)	−0.1889*** (0.0083)	−0.1943*** (0.0065)	−0.0024*** (0.0002)	−0.0023*** (0.0002)	−0.0023*** (0.0002)
age ²	0.0019*** (0.0001)	0.0018*** (0.0001)	0.0022*** (0.0001)	0.0018*** (0.0001)	0.0018*** (0.0001)	0.0022*** (0.0001)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)
schooling	0.0366*** (0.0051)	0.0311*** (0.0061)	0.0052 (0.0039)	0.0406*** (0.0052)	0.0461*** (0.0070)	0.0052 (0.0039)	0.0003*** (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)
pioneer	1.7622*** (0.0657)	1.5286*** (0.1024)	1.1741*** (0.0415)	1.8058*** (0.0670)	1.5418*** (0.1093)	1.1741*** (0.0415)	0.0713*** (0.0042)	0.0705*** (0.0041)	0.0692*** (0.0042)
distance_Ingdppc	0.8749*** (0.0888)	0.5826*** (0.0934)	0.2135*** (0.0376)	0.8861*** (0.0881)	0.5819*** (0.0948)	0.2135*** (0.0376)	0.0133*** (0.0032)	0.0086** (0.0041)	0.0107** (0.0045)
distance_coop	−0.0129 (0.0082)	0.0068 (0.0144)	−0.0073* (0.0039)	−0.0121 (0.0082)	0.0069 (0.0143)	−0.0073* (0.0039)	−0.0005** (0.0003)	−0.0004 (0.0003)	−0.0005 (0.0003)
distance_medical	46.4214* (27.9812)	43.7924 (42.9833)	26.8754** (12.4029)	45.4386 (27.9493)	45.3497 (43.6284)	26.8756** (12.4029)	2.5699*** (0.8988)	2.2949** (0.9026)	2.6314** (1.0154)
distance_highEdu	3.8617 (2.4409)	6.5296 (4.4566)	0.4295 (0.7884)	3.9861 (2.4537)	6.7124 (4.4735)	0.4295 (0.7884)	0.1378 (0.0893)	0.2271** (0.0882)	0.2377** (0.0919)
distance_ppDen	0.0464 (0.0771)	0.1925** (0.0809)	−0.0033 (0.0205)	0.0475 (0.0773)	0.1953** (0.0819)	−0.0033 (0.0205)	0.0046 (0.0043)	0.0159*** (0.0050)	0.0160*** (0.0049)

distance_tertiaryRatio	0.2462 (0.5814)	1.8194** (0.7838)	-0.1408 (0.2204)	0.2182 (0.5774)	1.7686** (0.7913)	-0.1408 (0.2204)	-0.0269 (0.0300)	-0.0673** (0.0323)	-0.0847** (0.0347)
Constant	-1.8388*** (0.3101)	-1.5829*** (0.2573)	0.0083 (0.2468)	-1.9502*** (0.3137)	-1.7668*** (0.2857)	0.0083 (0.2468)	0.0612*** (0.0046)	0.0593*** (0.0046)	0.0593*** (0.0045)
Level 2 variance	0.1484 (0.0372)	1.6617 (0.1672)	1.2686 (0.1010)	0.1870 (0.0312)	0.2416 (0.0455)	0.0000 (0.0000)			
Level 3 variance			0.1054 (0.0398)	1.6046 (0.1685)	1.2686 (0.1010)				
ICC	0.0432 (0.0104)	0.3356 (0.0224)	0.2783 (0.0160)	0.0816 (0.0112)	0.3595 (0.0214)	0.2783 (0.0160)			
Time FE	Y	Y	Y	Y	Y		Y	Y	Y
Industry FE							Y	Y	Y
Num. of instruments							182	180	201
AR(2)							0.772	0.831	0.824
Hansen's J test							0.368	0.278	0.408
<i>Difference-in-Hansen tests</i>									
GMM instruments for levels – Excluding group							0.123	0.103	0.113
GMM instruments for levels – Difference (null H = exogenous)							0.466	0.377	0.622
GMM instrument for distance_trend – Excluding group							0.553	0.381	0.521
GMM instrument for distance_trend – Difference (null H = exogenous)							0.103	0.179	0.185
Obs	749,219	749,219	749,219	749,219	749,219	749,219	729,965	729,965	729,965

Notes: Trending indicators estimated here are untransformed, i.e. $E_Trending_{ik,t} - E_Trending_{ij,t}$. Standard errors shown in parentheses are clustered at the destination city, except for multilevel logit models, where robust standard errors are applied but not clustered. The Windmeijer correction (Windmeijer2005) is enabled in the GMM estimation. Coefficients are displayed as 0.0000 because the values are smaller than 0.0001. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: Created by authors using CHFS and China Data Institute (2021).

Appendix 2 Regression tables of complementary results

Table A3. Determinants of migration decisions (1997–2017): employment growth rates.

	OLS		Multilateral resistance to migration			
	(1)	(2)	(3)	(4)	(5)	(6)
distance_EGrowth	0.0035*** (0.0008)	0.0006 (0.0005)	0.0002 (0.0004)	0.0006 (0.0003)	0.0005 (0.0006)	0.0001 (0.0005)
gender		0.0009*** (0.0001)	0.0006*** (0.0001)	0.0001 (0.0001)	0.0006*** (0.0001)	0.0006*** (0.0001)
marriage		−0.0079*** (0.0006)	−0.0074*** (0.0005)	−0.0041*** (0.0005)	−0.0072*** (0.0005)	−0.0074*** (0.0005)
hukou_type		−0.0010** (0.0004)	−0.0004* (0.0003)	−0.0000 (0.0002)	−0.0004* (0.0003)	−0.0004* (0.0003)
health_status		−0.0002** (0.0001)	−0.0001* (0.0001)	−0.0001* (0.0001)	−0.0001* (0.0001)	−0.0001 (0.0001)
hh_income		0.0000 (0.0001)	−0.0000 (0.0001)	−0.0001** (0.0001)	0.0000 (0.0001)	−0.0000 (0.0001)
age		−0.0024*** (0.0002)	−0.0023*** (0.0002)	−0.0014*** (0.0002)	−0.0023*** (0.0002)	−0.0023*** (0.0002)
age ²		0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)
schooling		0.0002*** (0.0000)	0.0002*** (0.0000)	0.0000 (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)
pioneer		0.0670*** (0.0044)	0.0622*** (0.0036)	0.0667*** (0.0046)	0.0600*** (0.0034)	0.0617*** (0.0036)
distance_lngdppc		0.0141*** (0.0023)	−0.0011 (0.0045)	0.0023 (0.0048)	−0.0017 (0.0055)	−0.0031 (0.0045)
distance_coop		−0.0004 (0.0003)	−0.0007** (0.0003)	−0.0007** (0.0003)	−0.0008** (0.0004)	−0.0009*** (0.0003)
distance_medical		2.2484** (0.9989)	3.2036*** (0.9130)	2.9437*** (0.7792)	4.0403*** (1.2303)	3.2219*** (1.0067)
distance_highEdu		0.1465* (0.0815)	0.1033 (0.0940)	0.0485 (0.0885)	0.1771* (0.0964)	0.1322 (0.0957)
distance_ppDen		0.0025 (0.0019)	−0.0020 (0.0083)	0.0005 (0.0073)	−0.0099 (0.0075)	−0.0008 (0.0079)
distance_tertiaryRatio		0.0034 (0.0188)	−0.0374 (0.0264)	−0.0410 (0.0261)	−0.0291 (0.0272)	−0.0365 (0.0264)
Constant	0.0088*** (0.0009)	0.0616*** (0.0045)	0.0600*** (0.0042)	0.0394*** (0.0042)	0.0601*** (0.0042)	0.0597*** (0.0042)
Time FE	Y	Y	Y	Y		
Industry FE			Y	Y	Y	Y
Origin FE			Y		Y	
Destination FE			Y			Y
Pairs of cities FE				Y		
Origin-year FE						Y
Dest-year FE					Y	
R ²	0.0016	0.0458	0.0534	0.0840	0.0722	0.0698
Obs	1,038,407	758,754	739,308	739,303	739,109	739,067

Notes: The variable ‘distance_EGrowth’ is defined as $GR_{i,k,t} - GR_{ij,t}$. Standard errors shown in parentheses are clustered at the destination city. Coefficients are displayed as 0.0000 because the values are smaller than 0.0001.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: Created by authors using CHFS and China Data Institute (2021).

Table A4. Determinants of migration decisions (1997–2017): wage prospects with fixed effects.

	OLS		Multilateral resistance to migration					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
distance_wage	0.0237*** (0.0024)	0.0041*** (0.0014)	0.0039*** (0.0013)	0.0033** (0.0014)	0.0002 (0.0013)	0.0004 (0.0013)	−0.0000 (0.0015)	0.0002 (0.0013)
gender		0.0003*** (0.0001)	0.0006*** (0.0001)	0.0000 (0.0001)	0.0002** (0.0001)	−0.0001 (0.0001)	0.0003** (0.0001)	0.0002** (0.0001)
marriage		−0.0041*** (0.0005)	−0.0043*** (0.0005)	−0.0039*** (0.0005)	−0.0040*** (0.0004)	−0.0023*** (0.0003)	−0.0040*** (0.0004)	−0.0040*** (0.0004)
hukou_type		−0.0005 (0.0003)	−0.0005** (0.0002)	−0.0002 (0.0003)	−0.0001 (0.0002)	0.0001 (0.0001)	−0.0002 (0.0002)	−0.0001 (0.0002)
health_status		−0.0002* (0.0001)	−0.0001 (0.0001)	−0.0001 (0.0001)	−0.0001 (0.0001)	−0.0001 (0.0001)	−0.0001 (0.0001)	−0.0001 (0.0001)
hh_income		0.0002* (0.0001)	0.0002*** (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)	−0.0001 (0.0001)	0.0001* (0.0001)	0.0001* (0.0001)
age		−0.0021*** (0.0002)	−0.0020*** (0.0002)	−0.0020*** (0.0002)	−0.0019*** (0.0002)	−0.0010*** (0.0001)	−0.0019*** (0.0001)	−0.0019*** (0.0001)
age ²		0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)	0.0000*** (0.0000)
schooling		0.0002*** (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)	0.0000 (0.0000)	0.0002*** (0.0000)	0.0002*** (0.0000)
pioneer		0.0785*** (0.0046)	0.0794*** (0.0044)	0.0737*** (0.0041)	0.0744*** (0.0040)	0.0778*** (0.0052)	0.0735*** (0.0039)	0.0743*** (0.0041)
distance_unemploy		−0.0845 (0.0970)	−0.0766 (0.0926)	−0.0808 (0.0979)	−0.1569 (0.1124)	−0.1479 (0.1154)	−0.1943* (0.1054)	−0.2129 (0.1310)
distance_coop		0.0002 (0.0003)	0.0001 (0.0002)	0.0002 (0.0003)	−0.0006 (0.0004)	−0.0006 (0.0004)	−0.0005 (0.0005)	−0.0006* (0.0003)
distance_medical		3.1689*** (0.8900)	3.3611*** (0.8360)	3.0711*** (1.0386)	1.6373** (0.6755)	1.2017 (0.7307)	2.2222** (0.9063)	1.4692** (0.7346)
distance_highEdu		0.2206*** (0.0689)	0.1963*** (0.0632)	0.2220*** (0.0751)	0.0491 (0.0632)	−0.0043 (0.0626)	0.0803 (0.0780)	0.1016* (0.0614)
distance_ppDen		0.0042** (0.0018)	0.0036* (0.0018)	0.0056*** (0.0019)	−0.0037 (0.0080)	−0.0041 (0.0082)	−0.0114** (0.0051)	−0.0028 (0.0078)

(continued).

Table A4. Continued.

	OLS		Multilateral resistance to migration					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
distance_tertiaryRatio		0.0070 (0.0169)	−0.0046 (0.0157)	0.0218 (0.0173)	−0.0162 (0.0201)	−0.0256 (0.0217)	−0.0027 (0.0205)	−0.0135 (0.0219)
Constant	0.0074*** (0.0007)	0.0478*** (0.0034)	0.0470*** (0.0034)	0.0457*** (0.0033)	0.0460*** (0.0030)	0.0293*** (0.0031)	0.0458*** (0.0030)	0.0456*** (0.0031)
Time FE	Y	Y	Y	Y	Y	Y		
Industry FE			Y	Y	Y	Y	Y	Y
Origin FE			Y		Y		Y	
Destination FE				Y	Y			Y
Pairs of cities FE						Y		
Origin-year FE								Y
Dest-year FE							Y	
R ²	0.0072	0.0418	0.0442	0.0460	0.0501	0.0851	0.0671	0.0683
Obs	1,050,567	664,285	664,284	664,283	664,282	664,212	664,121	664,074

Notes: Standard errors shown in parentheses are clustered at the destination city. Coefficients are displayed as 0.0000 because the values are smaller than 0.0001. The small R^2 is endemic to and a direct result of our discrete choice setting. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: Created by authors using CHFS, CHRI and China Data Institute (2021).

Appendix 3 Additional figures

The number of migrants moving out

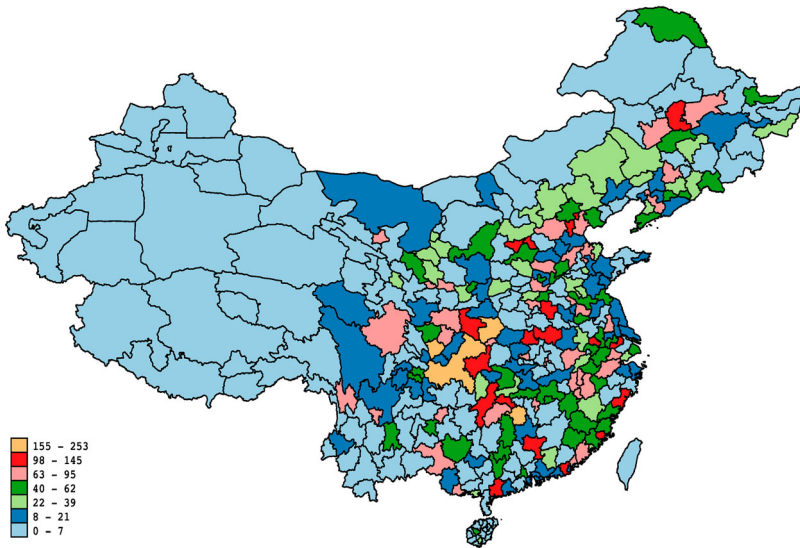


Figure A1. Geographic distribution of emigrants across Chinese prefecture-level cities.

Author's elaboration using CHFS.

Appendix 4 Data wrangling report

Introduction

The China Household Financial Survey was conducted biannually between 2011 and 2017. The latest wave tracked the majority of preceding households plus newly-surveyed families where totally 127,012 individuals were involved.³⁹ Each wave has three sets of data providing information on individuals, households, and cities. We used the latest wave because the information on individuals' origin cities are absent in earlier waves, and it satisfies all the features our approach needs: (a) it includes both natives and migrants, and (b) enables us to compile origin-destination information at the prefecture city level.

It should be noted that, as a household survey, answers in relation to family members were made by the respondent who acted as a delegate of his or her family. The CHFS's interviewers chose the person who was most familiar with his or her household's economic conditions, though this might not be feasible sometimes. In other words, the respondent can either represent him- or herself or share certain family relationships with the person who was indicated (parent, spouse, etc.). This design entails a few concerns. For example, when the respondents were on their own behalf, their resident cities are not clearly stated in the data since the question only asked where their family members were living now, if they did not live together. Instead, we can learn in which city respondents got surveyed from the city table. Yet where people got surveyed are not always the city where they mostly lived. In two ways, we clearly identified this discrepancy. Literally, as the 2017 questionnaire asked if the current city/county was the place where the family's main economic activities were carried out, an individual who did not stay with family members but visited or went back when the survey was conducted can be a respondent, although, intuitively, they should not be. While technically, two variables help us check if individuals lived in their Hukou registration cities/counties/towns. Based on them, we can make a comparison between the Hukou registration city and surveyed city. A few thousands of observations are found dubious. For instance, the variable A2019b shows that an individual lived in his or her Hukou registration city but he or she was surveyed in another city. To clean and compile the data, when surveyed cities are used as resident cities, we only retained householder samples.

Migrant identification

Overall, there are four types of migrants. First, based on the variable A2019e, we distinguished the floating population from natives and migrants with Hukou transfer.⁴⁰ The 2017 questionnaire annotates that the referring question would only be delivered if individuals' Hukou registration cities are different from their resident cities. As shown in Figure A2, we specified A2019e to be not null. Also, sometimes the data is messy, displaying information dissimilar to what we are told by the questionnaire, as shown in Figure A3. To avoid mistreating certain migrants as natives resulting from missing inputs in A2019e, we coded a supplementary condition starting by transfer = 0.⁴¹ Another type that we can straightforwardly identify is returnees. When the CHFS system detected that the individual's Hukou registration city is identical to his/her resident city, the interviewer then moved on to Question A2023g – if he/she had ever left the Hukou registration city for somewhere else to for more than six months. It is worth noting that although interviewers only asked family members who were aged above 16 in 2017, returnees can be quite young when they migrated. In our final dataset, these observations were dropped.

Identifying another two types is more demanding. By looking at the variable A2019e, we can learn whether the individual was living in the Hukou registration city or not, but there were some migrants with Hukou transfer. To avoid overestimating natives, we need other variables to help us find these samples, namely, A2022k (called 'transfer' in codes) and A2022m. The codes perhaps look a bit lengthy but, basically, we found four cases – migrant who lived in the county where his/her new Hukou was

```
case
  when a2019e is not null then 1
  when a2023g = 1 then 1
  when
    transfer = 0 and ((a2001 != 1 and a2016b_short != a2019_short) or
      (a2001 = 1 and hhead = 1 and a2016b_short != a2019_short)) then
      1
  when transfer = 1 and (((a2019_short != a2022m_short and a2019b is not null) or
    (a2019b is null and
      ((a2016b is null and surveyedCity != a2022m_short and
        (a2001 != 1 or (a2001 = 1 and hhead = 1)))
      or (a2016b is not null and a2016b_short != a2022m_short and
        (a2001 != 1 or (a2001 = 1 and hhead = 1)))))) then 1
  else 0 end migrant,
```

Figure A2. SQL codes for identifying migrants.

290 rows							
	hhid	pline	a2001	hhead	a2016b	a2019	a2019e
1	201103874	1	7	1	31-3101-310101	34-3412-341226	<null>
2	2015021551	3	6	<null>	52-5201-520102	41-4113-411323	<null>
3	2017000201	6	6	<null>	33-3300-330000	42-4210-421023	<null>
4	2017000415	2	2	<null>	31-3101-310107	32-3208-320803	<null>
5	2013013344	6	6	<null>	35-3506-350602	35-3507-350702	<null>
6	2013013344	5	6	<null>	35-3501-350102	35-3507-350702	<null>
7	2015035622	3	6	<null>	43-4300-430000	62-6201-620104	<null>
8	2015018175	5	2	1	12-1200-120000	35-3509-350923	<null>
9	2013020837	5	6	<null>	44-4420-442000	45-4504-450406	<null>
10	2013020837	3	6	<null>	44-4420-442000	45-4504-450406	<null>
11	201107732	4	6	<null>	11-1100-110000	51-5116-511603	<null>
12	2015020005	1	10	<null>	44-4419-441900	44-4452-445201	<null>
13	2015032364	5	8	<null>	32-3205-320500	51-5115-511521	<null>
14	2017001638	5	7	<null>	32-3202-320201	52-5212-521203	<null>
15	2017001853	6	6	<null>	44-4401-440100	46-4604-460400	<null>
16	201103848	1	6	1	31-3101-310113	34-3412-341226	<null>
17	201105086	12	7777	<null>	53-5325-532522	41-4113-411302	<null>

Figure A3. Respondents whose Hukou registration cities and resident cities are mismatching.

Notes: These are failed to be identified by A2019e.



	hhid	pline	a2003	a2016b_short	a2019_short	a2019b	transfer	a2022e_short	surveyedCity
1	2015014195	2	2	<null>	3101	<null>	1	3302	3302
2	2017000273	2	2	<null>	3415	<null>	1	5107	3101
3	2017000301	2	2	<null>	4414	<null>	1	1408	4401
4	2017000355	2	2	<null>	5109	<null>	1	5002	5001
5	2015013599	4	7	<null>	3311	<null>	1	3301	3301
6	2017000356	2	2	<null>	4415	<null>	1	4416	4403
7	2017000490	2	10	<null>	6104	<null>	1	6110	6101
8	2011003103	5	6	<null>	6101	<null>	1	6103	6103
9	2017000512	1	1	<null>	3102	<null>	1	6529	3101
10	2017000510	2	2	<null>	1107	<null>	1	1510	1504
11	2017001086	5	6	<null>	4403	<null>	1	4409	4409
12	2015035430	2	2	<null>	6211	<null>	1	6201	6201
13	20130119521	4	6	<null>	4401	<null>	1	4408	4408
14	2017001293	4	6	<null>	4406	<null>	1	4401	4401
15	2017001429	2	2	<null>	3505	<null>	1	3505	3505
16	2015022596	3	2	<null>	4406	<null>	1	4401	4401
17	2017000842	3	7	<null>	5107	<null>	1	<null>	5101
18	2017001293	2	2	<null>	4406	<null>	1	4401	4401
19	2017001441	1	1	<null>	4403	<null>	1	2201	2102

Figure A4. Migrants with Hukou transfer: Surveyed cities are identical to origin cities.

registered; migrant who lived outside the Hukou registration county but inside the newly-registered prefecture city; migrant who lived neither in the newly-registered prefecture city nor in the origin city; migrant whose Hukou had been transferred to another prefecture city remained staying in his/her origin city (some examples are given in Figure A4).⁴² To exclude samples of the last case that are not captured by A2019e, we used the variable 'surveyed city' followed by A2001.⁴³ Lastly, a few people had migrated several times. The situation can be rather intricate – they transferred the Hukous after they migrated to new prefecture cities, yet the cities where their Hukous were newly registered are not the ones paired with the variables telling in which year they migrated. We specified these samples as migrants first and then dropped some of them if, after compiling, their destination cities and moving-in years are still mismatching.⁴⁴

Information pairwise

Now we successfully sorted out migrant observations, while the next issue lying ahead of us is to identify to which prefecture city that each migrant moved, and accordingly, in which year they moved out and moved in. Foremost, variables A2016b, A2019, A2023j and 'surveyed city' all provide usable information. To apply the correct one, we need to further classify migrants into different situations. As always, we endeavoured to improve the data accuracy and retain as many observations as possible.

In general, the classification can be summarised as follows:

- When A2023g equals 2, it suggests that individuals had never resided in other prefecture cities. While looking at migrants with Hukou transfer, sometimes the reference point that respondents took is the cities where their current Hukous were registered, instead of their 'real' origin cities. Therefore, by setting transfer equals 1 and A2023g equals 2 followed by A2016b, we assigned either current Hukou registration cities or resident cities to those samples as their destination, in terms of their specific situations.
- When A2023g equals 1, it simply means that individuals were returnees. In general, A2023j tells us where individuals had migrated to before they returned to the Hukou registration cities. Yet for certain observations, this information is missing. Instead, we assigned current Hukou registration cities as destination to some of them if they are migrants with Hukou transfer.
- When A2023g is null, the CHFS annotates that this question was not assigned to those individuals because they lived in the prefecture cities that their Hukous did not pertain to. In this case, we may learn about their resident cities using A2016b. However, sometimes even though we know in which city the migrant was living, his/her moving-out and/or moving-in year, referring to A2019e and A2019f, is still missing, and thus, these observations cannot be used. As an alternative solution to keeping information pairwise, we assigned the new Hukou registration cities to them and re-calculated their migration years by assuming they moved one year before the time that they transferred the Hukous.⁴⁵
- Lastly, among the floating population, destination cities can be either A2016b or 'surveyed city' depending on if A2016 has a valid value. Again, we only applied the variable 'surveyed city' to householders (hhead = 1).

hhid	pline	a2001	a2016b_short	a2019_short	a2019b	transfer	a2022m_short	surveyedCity
1 2017000273	2	2	<null>	3415	<null>	1	5107	3101
2 2017000301	2	2	<null>	4414	<null>	1	1408	4401
3 2017000355	2	2	<null>	5105	<null>	1	5002	5001
4 2017000356	2	2	<null>	4415	<null>	1	4416	4403
5 2017000499	2	10	<null>	6104	<null>	1	6110	6101
6 2017000512	1	1	<null>	3102	<null>	1	6529	3101
7 2017000530	2	2	<null>	1307	<null>	1	1310	1306
8 2017001441	1	1	<null>	4403	<null>	1	2201	2102
9 2017023822	2	10	<null>	3501	<null>	1	1101	3101
10 2017061173	1	1	<null>	2201	<null>	1	2205	2202
11 2017029472	5	5	<null>	1401	<null>	1	1401	1101
12 2017029472	3	3	<null>	1401	<null>	1	1401	1101
13 2015016963	2	2	<null>	3506	<null>	1	3501	3502
14 2013021805	2	2	<null>	4201	<null>	1	4303	5001
15 2015026797	2	2	<null>	3607	<null>	1	3601	4403
16 2017029472	4	5	<null>	1401	<null>	1	1401	1101
17 2017018396	2	2	<null>	3213	<null>	1	3209	3101
18 2013004508	1	2	<null>	1509	<null>	1	1307	1502
19 2015000846	5	10	<null>	4104	<null>	1	6542	1101

Figure A5. Migrants with Hukou transfer: New and old Hukou registration cities as well as surveyed city are all non-identical.

```

case when migrant=1 and transfer=1 and a2023g=2 and a2016b is null and a2019_short!
=a2022m_short then a2019_short
  when migrant=1 and transfer=1 and a2023g=2 and a2016b is not null and a2016_revised!
=a2022m_short then a2016_revised
  when migrant=1 and transfer=1 and a2023g=1 and a2023j is not null then a2023j
  when migrant=1 and transfer=1 and a2023g=1 and a2023j is null and a2019_short!
=a2022m_short then a2019_short
  when migrant=1 and transfer=1 and a2023g is null and a2019e or a2019f is not null then
a2016_revised
  when migrant=1 and transfer=1 and a2023g is null and (a2019e and a2019f) is null and
a2019_short!=a2022m_short then a2019_short
  when migrant=1 and transfer=0 and a2023g is null and a2016b is null and hhead=1 and
a2001=1 and surveyedCity!=substring(origin,4,4) then surveyedCity
  when migrant=1 and transfer=0 and a2023g is null and a2016b is not null then
a2016_revised
  when migrant=1 and transfer=0 and a2023g=1 then a2023j else substring(origin,4,4)
end destination

```

Figure A6. SQL codes for identifying migration destination.

Codes used for compiling the destination are summarised in Figure A6. Besides the aforementioned, we assigned the origin cities (A2022m) to all natives as their destination. Moreover, as shown in Figure A7, sometimes the moving-out year (A2019e) is not the same as the moving-in year (A2019f) because individuals moved more than once. As they are not useful in our analysis, we dropped these observations.

Final cleaning procedure

Migrant samples. Afterwards, there is still a bit of information not pleasing us because some observations' inputs remained strange. For instance, the destination city does not necessarily match the moving-in year, and a small cohort of respondents did not answer questions based on the common knowledge, as shown in Figure A8. First, dozens of samples present available inputs in A2023j, but the corresponding data that should be stored in A2023k is missing. Hence, the destination variable extracted the data from A2023j but used the proxy (the year of Hukou transfer minus 1) as the migration time. To fix this problem, we can either replace the destination variable with A2019 (the current Hukou registration city) or simply drop them. We chose the latter. Moreover, some observations have identical inputs in A2022m and A2023j. This could be explained by three possible reasons – the respondent took his/her current Hukou registration city as the reference point and thus considered where he/she came from as the place he/she had stayed in outside the Hukou registration city; individuals migrated and transferred the Hukous more than once; individuals returned to their origin cities even after they transferred the



a2019b	a2019e	a2019f	transfer	a2022n_short	a2022l	a2023k	surveyedCity	leaveTime	arriveTime
<null>	<null>	<null>	1	4107	2016	<null>	5101	2015	2015
<null>	2001	2012	1	3301	2001	<null>	6101	2001	2012
<null>	1991	1991	1	5107	1980	<null>	3101	<null>	<null>
<null>	<null>	<null>	1	1522	1979	1980	1201	<null>	<null>
2007	2007	<null>	1	1408	2014	<null>	4401	2007	2007
2013	2013	<null>	1	5326	2009	<null>	5323	2013	2013
2016	2016	<null>	1	5002	1966	<null>	5001	2016	2016
2007	2007	<null>	1	4416	2008	<null>	4403	2007	2007
<null>	2007	<null>	1	6110	2012	<null>	6101	2011	2007
<null>	<null>	<null>	1	6529	1964	1964	3101	<null>	<null>
<null>	<null>	<null>	1	4212	2008	<null>	3502	2007	2007
2009	2014	<null>	1	1310	2009	<null>	1306	2009	2014
<null>	<null>	<null>	1	3206	2015	<null>	3101	2014	2014
2014	2014	<null>	1	1401	2015	<null>	4331	2014	2014
<null>	<null>	<null>	1	1304	1999	<null>	1301	1998	1998
<null>	<null>	<null>	1	6109	1973	1978	6101	<null>	<null>
<null>	<null>	<null>	1	3611	1981	1971	3301	<null>	<null>
2015	2015	<null>	1	2310	2002	<null>	3101	2015	2015
2017	2017	<null>	1	3410	2017	<null>	3412	2017	2017

Figure A7. Example of migration year.

a2016...	a2019...	a2019e	a2019f	tra...	a2022...	a2022l	a2023j	a2023k	leaveTime	arriveTime	destination
3205	3205	<null>	<null>	1	2303	2006	3717	<null>	2005	2005	3717
3205	3205	<null>	<null>	1	3701	2006	<null>	<null>	2005	2005	3205
2112	2112	<null>	<null>	1	2203	1994	<null>	<null>	<null>	<null>	2112
<null>	5002	<null>	<null>	1	5309	2016	5309	1985	2015	2015	5309
<null>	2201	<null>	<null>	1	3101	2001	3101	1997	1997	1997	3101
<null>	2201	<null>	<null>	1	2224	2012	2224	1989	2011	2011	2224
<null>	2201	<null>	<null>	1	3210	2008	3210	2003	2003	2003	3210
5101	5101	<null>	<null>	1	5002	1987	4403	1997	1997	1997	4403
<null>	4401	<null>	<null>	1	2102	2007	<null>	<null>	2006	2006	4401
4420	4290	2010	2010	1	4201	2015	<null>	<null>	2010	2010	4420
<null>	5002	<null>	<null>	1	5101	1987	5101	1974	<null>	<null>	5101
5309	5309	<null>	<null>	1	4502	2017	<null>	<null>	2016	2016	5309
<null>	1202	<null>	<null>	1	1410	2016	1410	1991	2015	2015	1410
<null>	1102	<null>	<null>	1	1507	2008	1507	2008	2008	2008	1507
<null>	2201	<null>	<null>	1	2224	2003	2224	1987	2002	2002	2224
<null>	2201	<null>	<null>	1	2204	1980	2204	1978	<null>	<null>	2204
<null>	3701	<null>	<null>	1	3412	2015	3702	2014	2014	2014	3702
6106	4401	2016	2016	1	6207	2000	<null>	<null>	2016	2016	6106
<null>	2201	<null>	<null>	1	2224	1980	<null>	<null>	1997	1997	2201
3302	3508	2012	2012	1	3505	2005	<null>	<null>	2012	2012	3302

Figure A8. Remaining errors.

Hukous. By setting $A2022l < A2023k$, we can easily identify the last group, but it is hard to distinguish the first two. As a result, although we know they are migrants, we cannot identify in which year these migrants moved to their resident cities. Thus, these observations were eventually dropped as well.

Recall that certain returnees may migrate when they were quite young (under 16 years old). As we aim to examine labour migration, teenagers are not proper to be included. After merging the datasets, we transformed the age variable to be time-variant. In this way, we eliminated all the observations if their ages were either below 16 or above 65 between 1997 and 2017. In addition, for migrants who worked previously but did not have a job when they were interviewed in 2017, we matched regional employment statistics for them based on the industry categories of their last employment. Concerning they may migrate after being unemployed or retired, only respondents whose last jobs' termination dates ($A3139$) are later than their migration time were kept in the data.

Native samples. Native samples are much simpler to clean. Hereby, we give a brief summary:

- Drop observations if $A2022k$ (transfer) equals 1 but $A2019$ is null.
- Drop observations if $A2022k$ (transfer) equals 1 but $A2019b$ and $A2023g$ are null.
- Drop observations if the city retrieved from the variable 'surveyed city' is non-identical to the origin city while the respondent was identified as a householder.

Lastly, among the full sample (both natives and migrants), observations were dropped if $A3138$ equals 2 (who had never worked).

Variables summary

Table A5. Variables – questions.

Variables	Questions
A2001	XXX is your? [Example: Myself; Spouse; Parents; Children]
A2016b	In which province/city/county does XXX live?
A2019	Which province/city/county is the registered residence of XXX?
A2019b [†]	Is XXX Hukou registered in the villages/towns where he/she now lives?
A2019e [‡]	In which year did XXX leave [A2019]?
A2019f [‡]	In which year did XXX come to his/her resident province/city?
A2022k	Has XXX ever transferred the Hukou to another district/county?
A2022l	In which year did XXX experience his/her latest Hukou transfer?
A2022m	XXX's Hukou is moved out from which county/city/province?
A2023g*	Has XXX ever left [A2019] for somewhere else to for over 6 months?
A2023j*	Which province/city did XXX live before returning?
A2023k*	In which year did XXX go to [A2023j]?
A3138	Has XXX worked before?
A3139	When did XXX's last job end?

Notes: [†] denotes questions that are asked only if the CAPI system detects that the individual's resident county/district is identical to his/her Hukou registration county/district. [‡] denotes questions that are delivered only if the CAPI system detects that the individual's resident prefecture city is non-identical to his/her Hukou registration city. * denotes questions that are asked only if the CAPI system detects that the resident prefecture city of an individuals, whose age has been above 16 by the time of survey, is identical to his/her current Hukou registration city.

Source: China Household Finance Survey (2017).