Received: 14 October 2020 | Revised: 5 April 2021 | Accepted: 27 May 2021

DOI: 10.1002/psp.2493

RESEARCH ARTICLE

WILEY

Determinants and consequences of rural-to-urban migration patterns in China: Evidence from sequence analysis

Zhenxiang Chen[1] | Yao Lu[2] | Donald J. Treiman[3]

[1]Department of Sociology, University of California, Los Angeles, Los Angeles, California, USA

[2]Department of Sociology, Columbia University, New York, New York, USA

[3]California Center for Population Research, University of California, Los Angeles, Los Angeles, California, USA

Correspondence
Zhenxiang Chen, Department of Sociology, University of California, Los Angeles (UCLA), Los Angeles, CA, USA.
Email: zchen716@ucla.edu

Funding information
National Institute of Child Health and Development, Grant/Award Number: K01HD073318; Eunice Kennedy Shriver National Institute of Child Health and Human Development, Grant/Award Number: R24HD041022; National Science Foundation, Grant/Award Number: 0551279

Abstract

This paper adopts a life course perspective that captures the migration trajectory for rural-to-urban migrants in China during the observation window. By taking this trajectory approach, we aim to advance the understanding of divergent rural-to-urban migration patterns in China and their determinants and consequences. We use data from the Survey of Internal Migration and Health in China (IMHC) and focus on individuals' migration experience between age 14 and 40. First, we apply sequence analysis to characterise the main rural-to-urban migration patterns by different timing, duration, frequency, and direction and identify seven common patterns. We next examine the determinants of these patterns. The results suggest that demographic characteristics, socio-economic background, and hometown characteristics shape migration trajectories in complex ways, highlighting that social origin can substantially determine migration patterns of rural Chinese. Furthermore, we examine via a counterfactual framework how the seven migration patterns shape migrants' occupational attainment while taking account of self-selection into different migration trajectories. The findings show that (i) there is self-selection into migration trajectories that has implications for occupational status and (ii) non-transient adult urban migration is associated with higher occupational attainment whereas other types of migration are not.

KEYWORDS
China, rural-to-urban migration, selection, sequence analysis, social mobility

1 | INTRODUCTION

Rapid economic development in China has led to massive migration from rural to urban areas. Rural-to-urban migration in China represents one of the most extensive human migration flows in the world. The number of rural-to-urban migrants reached 288 million in 2018 (National Bureau of Statistics of the People's Republic of China, 2019), a size comparable to the total global international migration. Although the intensity of migration is low (M. Bell et al., 2015; M. Bell et al., 2020), the sheer magnitude makes Chinese rural-to-urban migration an important phenomenon to study. Equally important, the complexity and heterogeneity (e.g., different timing, duration, frequency, and direction) of internal migration in China provide unique opportunities to uncover distinct migration patterns and their consequences.

Studies of migration patterns have mostly relied on cross-sectional data and have reached conclusions regarding migration patterns from measurements of migration status at one or a few points in time (but for recent exceptions see below). This conventional approach reduces complex migration patterns to isolated moves or counts of cumulative migration years. It leads to a static understanding of migration patterns, which obscures the complex and heterogeneous dynamics of migration over the life course.[1] Imagine a group of migrants who sequentially undertake rural migration, then urban migration, and finally return migration and another group of migrants who experience urban migration, then return migration, and finally

rural migration. The two groups are clearly characterised by distinct migration patterns but, depending on the point of observation, may be classified similarly as rural, urban, or return migrants.

In recent years, growing efforts have been made to overcome this limitation in the study of migration patterns by using longitudinal data (R. Coulter et al., 2011; R. Coulter et al., 2016; R. Coulter & Van Ham, 2013). Beyond the use of longitudinal data, studies have also turned to retrospective data. Some studies have applied cohort approaches to retrospective data to study migration patterns over the life course (A. Bernard, 2017; A. Bernard et al., 2019; Falkingham et al., 2016) or study how past migration experience affects future migration from a life course perspective (A. Bernard & Perales, 2021; A. Bernard & Vidal, 2020).

Another main approach to the study of migration patterns over the life course using retrospective data is sequence analysis, which takes a holistic view of migration by analysing each individual's migration trajectory within a particular age range. Sequence analysis permits identifying and describing typical migration patterns from highly complex and diverse migration trajectories over the (whole or partial) life course (A. Abbott & Forrest, 1986; Aisenbrey & Fasang, 2010; Brzinsky-Fay & Kohler, 2010). Although sequence analysis has increasingly been used in demographic research, its extension to the study of migration is limited. Stovel and Bolan (2004) has first applied sequence-based method to identify residential trajectories that describe patterns of movement across geographic landscapes and found that residential trajectories differ by the stage of adulthood. However, the use of sequence analysis on migration studies has begun to take off only recently. Our study joins the small but growing number of studies that use sequence analysis to examine migration patterns across substantial age ranges (di Belgiojoso & Terzera, 2018; Impicciatore & Panichella, 2019; Liao & Gan, 2020; Stovel & Bolan, 2004; Vidal & Lutz, 2018; M. Yang et al., 2020; Zufferey et al., 2021).

These studies vary in their focus and in the details of their implementation, which provide important insights upon which we build. In addition to differences in the country or region of the study, these studies importantly differ from each other in how they define migration states. These differences, both in locale and in methods, yield substantial differences in the number of potential migration trajectories. For example, interested in family migration trajectories for migrants in Italy, di Belgiojoso and Terzera (2018) define migration states based on family reunification status (alone, single, and reunited) and have identified various family migration trajectories (e.g., partial, slow, and quick family reunification trajectories) in a 10-year observation window. Similarly, to study multiple and repeat migration in Switzerland, Zufferey et al. (2021) define migration status by internal and international migration moves and have identified multiple migration trajectories (e.g., internal mobility and international circulation trajectories) in a 5-year observation window.

Beyond the identification of migration patterns, as variously defined, the research goal of these previous studies is often to explore the determinants and/or consequences of divergent migration patterns. We learn from these studies that there are selective processes in migration patterns and that exploring determinants of these patterns is important to the understanding of why and how migrants end up in different trajectories. For example, di Belgiojoso and Terzera (2018) show that family, cultural, and gender norms of the home country and the nature of migration (e.g., settlement vs. temporary) determine family migration trajectories. Focusing on internal migration patterns in West Germany, Vidal and Lutz (2018) find that internal migration is related to life-course transitions. Impicciatore and Panichella (2019) explore south-to-north internal migration trajectories in Italy and show that the characteristics of migrants in different migration trajectories differ significantly.

Some of the prior studies have gone beyond the determinants of migration trajectories to show that migration trajectories have significant consequences for migrants. For example, Impicciatore and Panichella (2019) show that different migration trajectories are associated with social mobility outcomes. M. Yang et al. (2020) study internal migration trajectories in China and find that these trajectories are associated with mental health. Liao and Gan (2020) explore international migration trajectories for Filipino and Indonesian female domestic workers in Hong Kong and find that more complex migration trajectories are associated with higher job satisfaction.

We differ from these existing studies in our focus on internal migration in China (except for M. Yang et al., 2020), and in the way, we define migration states in our sequence analysis.[2] We capture multiple migration states simultaneously characterised by the timing, duration, frequency, and direction of migration. More fully and accurately measuring migration sequences is key to understanding the dynamics of migration and its determinants and consequences. We examine both the determinants and consequences of migration patterns in China. For the latter, we measure occupational status as an outcome of the migration trajectory and take account of self-selection into each trajectory to more accurately assess the consequences of migration patterns.

Specifically, using detailed migration history from the Survey of Internal Migration and Health in China (IMHC), we (i) identify and describe common yet distinct migration trajectories among rural-to-urban migrants in China over their life course; (ii) assess the factors that determine different migration patterns; and (iii) examine how different migration patterns relate to occupational attainment (while adjusting for possible selection into different patterns).

The remainder of the paper is structured as follows. Section 2 reviews approaches to measure migration and then reviews migration patterns in China and their determinants and consequences. Section 3 introduces the data set and methods employed. This section also describes the migration patterns identified. Section 4 presents the determinants and consequences of migration patterns identified and includes some sensitivity analyses. Section 5 summarises our main findings and contributions and further discusses the limitations of our analysis.

2 | BACKGROUND

2.1 | The conventional approach versus the dynamic approach to measuring migration

Many migration studies rely on cross-sectional data, which necessitate a static measure of migration. The static view tends to understate the complexities of migration and conflates different groups of migrants. For example, those who are not migrants at the time of a survey may include those who have never migrated, those who have migrated but have temporarily returned, and those who have permanently returned.

Increasingly, research has resorted to longitudinal survey data to study migration patterns. This strategy presents a notable improvement over snapshot measures by examining migration statuses at multiple points in time (R. Coulter et al., 2011; R. Coulter et al., 2016; R. Coulter & Van Ham, 2013). Although longitudinal surveys could be subject to sample attrition, especially in migration studies because attrition is disproportionately high for migrants relative to the general population, this may not necessarily lead to wrong estimates (Alderman et al., 2001). Therefore, longitudinal surveys could be ideal to study migration patterns. However, longitudinal data sets that include detailed information on migration for a substantial range of ages are quite rare, and no such data sets exist for China.

A useful alternative to longitudinal surveys is the collection of retrospective migration histories. Two longitudinal data sets, the China Health and Retirement Study (CHARLS) and the China Labor Force Dynamics Survey (CLDS), include retrospective migration histories in a single wave (CHARLS) and in two waves 2 years apart (CLDS). However, our inspection of the CLDS data suggests that its migration data have many inconsistencies. CHARLS would be a plausible alternative to IMHC and would have the advantage of including a much larger sample, but it has several disqualifying limitations discussed in Appendix A.

Although retrospective reports may be vulnerable to recall bias, previous research shows that retrospective data can yield high quality information (Assaad et al., 2018; Beckett et al., 2001). Indeed, individuals can remember major life events experienced by themselves and family members, such as migration, with considerable accuracy (Smith & Thomas, 2003). Still, underestimation of migration is possible in retrospective migration histories (Schoumaker, 2014). Nevertheless, the overall quality of retrospective data depends on the particular survey design (e.g.,, using life-history grids can help respondents to recall past moves, as adopted in IMHC) (Belli, 1998; Blane, 1996). In the present research, we capitalise on high quality, detailed retrospective migration histories to investigate migration patterns over the multiyear observation window.

A life-course approach allows us to examine how long-term migration trajectories unfold over time by conceptualising migration as a continuous and serial process and by taking into account persistence, change, and heterogeneity in migration status (Mulder, 1993; Wingens et al., 2011). Migration is often categorised by its timing, duration, frequency, and direction. Specifically, migration can be categorised as early or late migration, depending on the age of first migration (Guven & Islam, 2015; Kimbro, 2009); short-term or long-term migration, depending on the duration of the stay away from home (Guilmoto, 1998; Tegegne & Penker, 2016); one-time or recurrent migration, depending on the number of migration trips (Constant & Zimmermann, 2011); or rural-bound, urban-bound, or return migration, depending on the direction (for internal migration; K.W. Chan, 2013). These categories are not exclusive, and migrants can experience multiple types of movement over the life course. The trajectory approach we take (via sequence analysis) allows us to differentiate migration patterns marked by different timing, duration, frequency, and direction.

In the literature, research on migration trajectories has largely focused on small-scale qualitative data (Favell, 2011; King, 2002). Our life-course approach is facilitated by sequence analysis, which can be applied systematically to large-scale data. Applying this analysis to migration allows us to move beyond pre-determined classifications of migration types and adopt a data driven approach to discover typical patterns characterised by temporally ordered sequences of migration events. We then examine questions about what factors shape these patterns and how such patterns may influence other social processes.

2.2 | Migration patterns in China and their determinants and consequences

Most previous migration studies in China focus on over-time patterns rather than life-course patterns (K.W. Chan, 2001; C.C. Fan, 2005; Z. Liang, 2001; Z. Liang, 2004; Shen, 2012). A number of studies have overcome this limitation by applying a cohort approach or a sequence analysis approach to retrospective data (A. Bernard et al., 2019; M. Yang et al., 2020). Considering the complexities of rural-to-urban migration in China, we expect that multiple distinct migration patterns can unfold over the life course.

Among different categorizations of migration, temporary versus permanent migration is a key dimension that captures migration patterns over the life course. In China, permanent rural-to-urban migrants are likely to have successfully changed their *hukou* status (X. Yang & Guo, 1999), often because they first migrated for educational purposes (X. Wu & Treiman, 2004). By contrast, temporary rural-to-urban migration is often initiated for work-related reasons and ends with return migration (Hu et al., 2011; Z. Liang, 2004; L. Meng & Zhao, 2018).

Migration can be further differentiated along four dimensions—timing, duration, frequency, and direction. Within the group of temporary or permanent migrants, the timing of migration may vary substantially, with some migrants undergoing the initial trip at younger ages than others. The former accumulate migration-related capital starting from a young age, whereas the latter tend to accumulate local capital before embarking on migration. Also, among temporary migrants, the duration of each trip and the overall length of stay at the destination tend to vary, with some staying for a longer term (S. Démurger & Xu, 2015; Wang & Fan, 2006). Similarly, the frequency

of migration is a differentiating factor: some migrants experience only one trip over their life course, whereas others have multiple episodes of migration (A. Bernard et al., 2019). Still, another distinction lies in the direction of migration, whether from a rural area to an urban destination, from one city onward to another city, from an urban area back to one's rural origin, and so on (A. Bernard et al., 2019; Hu et al., 2011).

The literature on the determinants of rural-to-urban migration in China is extensive (Cao et al., 2018; Chunyu et al., 2013; Z. Liang & White, 1996; Shen, 2012). However, previous research typically examined determinants of either the first migration or the migration status observed at the time of the interview. What is left unanswered is how early life conditions shape migration patterns over the life course.

Our analysis is informed by previous research on the determinants of rural-to-urban migration, which points to two main sets of determinants. The first set includes individual socio-demographic factors such as age, gender, marital status, number of children, and social networks, among others (Hu et al., 2011; X. Yang, 2000; X. Yang & Guo, 1999; Zhao, 2003). The second set includes structural factors such as the development gap between origin and destination places, land reallocation reform in the rural home, and opportunity structures in the urban labour market (A. Chen & Coulson, 2002; Z. Wu & Yao, 2003; Yan et al., 2014; K.H. Zhang & Song, 2003; Zhu, 2002). In our analysis of the determinants of life-course migration patterns, we focus on individual early-life socio-demographic factors and structural factors in the origin places (rather than contemporaneous factors) to avoid problems due to reverse causality. It is important to note that migration capital can also significantly affect migration behaviour. In particular, early migration experience can affect future migration (A. Bernard & Perales, 2021; De Jong, 2000). However, in the trajectory approach, this possibility is taken account of because past migration experience is incorporated as part of the migration pattern.

Previous studies of the consequences of rural-to-urban migration in China have investigated a host of outcomes. For example, compared with urban-to-urban migrants or urban residents, rural-to-urban migrants have a greater likelihood of remaining in low-skilled work (Y. Chen, 2011). Nevertheless, upward occupational mobility is possible for permanent rural-to-urban migrants (Ou & Kondo, 2013). Besides economic and occupational attainment, several studies have explored rural-to-urban migrants' health and subjective well-being and have shown that interpersonal and institutional discrimination, stigmatisation, and victimisation have detrimental effects on rural-to-urban migrants' physical and mental health (J. Chen, 2013; Cheung, 2013; X. Li et al., 2006). However, our knowledge regarding variations across different groups of migrants characterised by different timing, duration, frequency, and direction is very rudimentary at present.

In this paper, we provide a systematic investigation of differences in occupational attainment across migration groups. We focus on occupational attainment because it is one of the most critical aspects of social stratification and carries important implications for other realms of well-being. Also, a desire for a non-agricultural occupation, which is associated with higher income, is often a primary reason for rural-to-urban migration. Moreover, occupational attainment is generally more reliably measured than income or earnings, as the latter is vulnerable to reporting bias (Angel et al., 2019; Kim & Tamborini, 2014; Moore et al., 2000). Finally, our data provide occupational histories, which allow us to study how occupational attainment trajectories are associated with migration trajectories. In contrast to the analysis of determinants, where we consider only early-life conditions, the analysis of consequences incorporates more contemporaneous variables that can shape occupational attainment.

3 | DATA AND METHODS

3.1 | Data and sample

We use data from the Survey of Internal Migration and Health in China (IMHC). The IMHC was jointly conducted by the University of California, Los Angeles, and the Capital Medical University (Beijing) between November 2007 and May 2008 (https://ccpr.ucla.edu/IM-China/). It was designed to study the determinants, dynamics, and consequences of internal migration for health and well-being. The survey used a multistage stratified probability sampling approach to produce a nationally representative sample; see D.J. Treiman (2007) for details regarding the sample design. First, according to their level of migration, urbanisation, and the educational level of the population, about 50,000 township units in China were divided into 75 strata. Then, within each stratum, two township units were randomly selected with probability proportional to size; thus, 150 township units were the primary sampling units. Selected townships were further divided into small enumeration districts (EDs). For rural townships, these consisted of administrative villages. For densely settled rural townships or parts of townships and for urban 'streets' (the urban equivalent of townships), these were geographical units approximately 250 by 250 m, drawn insofar as was possible to use major roads as boarders. Within each township, four EDs were randomly selected. Within each ED, all households were listed and approached in random order. Five households were chosen at random as the primary sample, and additional households were chosen at random as a back-up sample. Adults within each household were then sampled in such a way as to produce completed interviews for five people in each ED and thus 20 interviews per township. The response rate was close to 70% for the analytic sample, which is comparable to other surveys conducted in China. The survey oversampled township units with large migration (in-migration or out-migration) to ensure a sufficient sample of migrants. The survey interviewed a total of 3,000 respondents aged 18 to 64 with a structured questionnaire. The data provide detailed information on respondents' demographic and socio-economic status, migration histories and characteristics, and health.

As noted earlier, IMHC represents one of two datasets for China with high-quality migration history information for a nationally representative sample. The dataset includes up to 24 episodes of migration.

The quality of the migration history is high in two main aspects—temporal and spatial. Fewer than 2% of the respondents have migration histories with any chronological errors, and fewer than 0.5% of the respondents have a gap or missing destination information in their migration history. The high data quality reflects both the survey design and the fieldwork protocol. To ensure data quality, detailed instructions were given to interviewers on how to complete the life history tables. The instructions included a series of consistency checks between information in the migration history table and between migration histories and other life histories included in the questionnaire. Our internal consistency checks suggest that the migration history reports are of reasonably high accuracy.

For the purpose of our study, we restrict our sample to rural-origin people, defined as those with agricultural *hukou* at age 14 (dropping 949 persons).[3] We focus on rural-to-urban migration, that is, migration trajectories of rural-origin people who had at least one episode of urban migration, for two reasons. First, rural-to-rural and urban-to-urban migration (dropping 430 persons) is smaller in scale and governed by different principles than rural-to-urban migration. In contrast, rural-to-urban migration is greater in magnitude and is theoretically more interesting because of the large disparities between places of origin and destination. Second, including rural-to-rural migrants does not change our substantive findings. When we include rural-to-rural migrants, they are clustered with those who migrated to rural areas shortly after a temporary stint in cities. Moreover, we restrict our sample to those who were at least 25 years old at the time of the survey (dropping 196 persons) because those who were younger generally did not have enough migration episodes to accurately observe a trajectory. We further drop people who migrated before age 14 because such migration usually represents tied family migration (313 persons). The final analytic sample size is 1,112.

3.2 | Methods and variables for studying migration patterns: Sequence analysis

We employ sequence analysis (A. Abbott & Forrest, 1986; Aisenbrey & Fasang, 2010; Brzinsky-Fay & Kohler, 2010) to examine migration trajectories over a portion of the life course, specifically between ages 14 and 40. The method is used to identify general patterns (A. Abbott & Hrycak, 1990) from a large number of specific trajectories. Sequence analysis is particularly suitable for studying trajectories characterised by polytomous discrete states such as migration statuses (e.g., urban migration, rural migration, and return migration).

To conduct the sequence analysis, we first define migration status at each age between ages 14 and 40. We begin at age 14 to avoid capturing tied family migration (e.g., children migrating with parents or moving to join parents or another relative). We choose age 40 as the end of the observation period to avoid having too many censored sequences (i.e., sequences with missing states at older ages for younger respondents) and to make most sequences of roughly equal length. For example, if age 50 or 60 were set as the age limit, we would have 12% or 16% censored sequences, respectively. Choosing 40 as the age limit also takes account of the fact that most migration episodes occur before this age (A. Bernard et al., 2014; Horowitz & Entwisle, 2018). Studying subsequences extended to include ages where few events occur can artificially increase the similarity between sequences and produce bias in the cluster analysis. Our age cutoff means that we essentially capture early-to-midlife migration trajectories. Because the observation window is between ages 14 and 40 and that we restrict our sample to those who were at least 25 years old at the time of the survey, right censoring is present for people younger than age 40 at the time of the survey. These individuals could have more migration episodes between their age at the time of the survey and when they reach 40 years of age. To assess the extent to which this right censoring may bias our identification of migration patterns, we conduct a sensitivity analysis restricted to migrants who were at least age 40. The results are discussed in Appendix D. We observe the exact same seven migration patterns as the main analysis. The fact that the restriction to age 40 yields essentially the same results suggests that right censoring is not a serious issue.

In order to carry out the sequence analysis, we need to arrive at a substantively meaningful way of describing the moves each individual makes, including the possibility of never moving. The task of the sequence analysis is then to determine whether the sequence of moves for individuals can be grouped into a small number of patterns and, if so, to describe these patterns. Given that we are interested in timing, duration, frequency, and direction of migration, we settle on eight migration states to characterise each move—precisely, seven types of moves plus never moving. These (1) remain at the rural place of origin—one's hometown[4]; (2) migrate to the first urban destination; (3) migrate to the second urban destination; (4) migrate to the third or higher-order urban destination; (5) migrate to the first rural destination; (6) migrate to the second rural destination; (7) migrate to the third or higher-order rural destination; and (8) return home. Although we have excluded those who experience only rural-to-rural migration, we cannot exclude rural-to-urban migrants who have also experienced rural-to-rural migration at some points of their migration. Therefore, it is important to include states (5)–(7). This variable is constructed from the migration history table. Table A1 shows the summary statistics of migration frequency by type of migration (i.e., rural, urban, and return).

First, we use information on the size of the destination for each trip to distinguish rural versus urban destinations. 'Small village (<1,000)', 'ordinary village (1,000–2,500)', 'large village (>2,500)', and 'township seat of a *xiang*' (the township seat is the administrative centre, i.e.,, the seat of government, of a *xiang*; a *xiang* is a rural township) are categorised as rural destinations, following the National Bureau of Statistics of the People's Republic of China (2008). 'Township seat of a *zhen*' (a *zhen* is a township with at least one town), 'county seat (the administrative centre, i.e., the seat of government of a county)', 'county-level city', 'prefecture-level city', 'province capital', and 'province-level city' are categorised as urban destinations. To identify return migration, we determine if the destination of a trip is the migrant's hometown, which is directly available in the data.

Second, we identify the frequency of migration using the detailed migration histories. For migrants who made multiple trips, we identify the rural versus urban status of each destination. We also distinguish first, second, or third or more urban trips and similarly for rural trips. We cap the frequency count at three because most migrants had three or fewer urban destinations (95.5%) and three or fewer rural destinations (99.9%). Third, we construct a file of person-age data that follows each person from ages 14 to 40 to capture the timing of initial migration and the duration of each trip.

The sequence analysis proceeds in three steps, detailed in Appendix B. Combining Ward's hierarchical fusion algorithm and the Partitioning Around Medoids (PAM) algorithm, we reach a seven-cluster solution that has a better fit than other solutions. This solution is also theoretically meaningful and has a sufficient number of observations in each cluster. We use TraMineR and the WeightedCluster package in R to conduct the sequence analysis (Gabadinho et al., 2011; Studer, 2013). We identify seven distinct yet common migration trajectories based on timing, frequency, duration, and direction. These clusters include, in descending order of frequency, (i) one-step early adult rural-to-urban migration; (ii) two-step early adult rural-to-urban migration; (iii) adolescent rural-to-urban migration; (iv) middle adult rural-to-urban migration; (v) rural migration; (vi) return migration; and (vii) transient rural-to-urban migration (discussed in detail below).

3.3 | Describing migration patterns

Figure 1 displays the sequence index plots. Each horizontal line represents an individual sequence over time, with age in the *x* axis, from ages 14 to 40. We use colours to represent different migration states.

The sequence analysis identifies seven substantively distinct clusters of trajectories that differ in timing, duration, frequency, and direction. Most migrants in Cluster 1 took their first urban migration trip between ages 14 and their late 20s and stayed in their first urban destination until age 40 or the time of the survey. Migrants in Cluster 2 first moved to urban areas during their late teens or early adulthood, stayed in their first urban destination for a short period of time, and then moved on to another city. Migrants in Cluster 3 are similar to those in Cluster 1 as they migrated to one urban destination and stayed there until age 40 or the time of the survey. However, they migrated during adolescence or early adulthood. Distinct from migrants in the first three clusters, most of the migrants in Cluster 4 migrated after age 30 and remained in the city to which they had moved until age 40 or the time of the survey. Those in Clusters 5 and 6 migrated as adolescents but differed from the early-first-migration clusters—Clusters 2 and 3. Migrants in Cluster 5 stayed in an urban destination for a very short period of time and then moved again to settle in a rural area other than their hometown. Migrants in Cluster 6, in contrast, stayed a few years in their first urban destination and returned to their hometown; a small share moved out again but the majority stayed in their hometown until the time of the interview. Finally, migrants in Cluster 7 represented the group with the greatest frequency of migration: they initiated migration during adolescence and moved to three or more urban destinations, spending a short period of time in each destination.

The sequence analysis results reveal several general migration patterns in China. First, a large share (around 68%) of rural-to-urban migrants first migrated during adolescence or early adulthood (Clusters 1, 2, 3, 6, and 7). Second, although many migrants remained in one urban destination during the observation window (Clusters 1, 3, and 4), the timing of migration diverged substantially. Some migrated during adolescence (Cluster 3), whereas others migrated mainly during early adulthood (Cluster 1) or middle adulthood (Cluster 4). Third, migrants who moved to a rural area other than their hometown after urban migration (Cluster 5) or who returned to their hometown (Cluster 6) tended to have a short urban stint. Finally, about a quarter (24%) of migrants participated in serial migration by moving to two or more urban destinations (Clusters 2 and 7). We notice that the last three clusters have a relatively small sample size. We also conduct sensitivity analyses with reduced number of clusters by combining some small clusters. The results are discussed in Appendix D. The sensitivity analyses yield largely similar results and conclusions.

The accuracy of sequence index plots decreases with sample size because multiple cases are plotted on top of each other ('overplotting'; Fasang & Liao, 2014), which means that individual sequences shown in the plot may no longer represent individuals but rather multiple individuals. One way to overcome this problem is to present state distribution plots (Billari & Piccarreta, 2005). Figure 2 does this, showing the percentage distribution of each migration state at each time point. These plots more clearly illustrate the migration trajectories detailed in the sequence index plots. Migrants in Cluster 1 ('One-Step Early Adult Urban') and Cluster 2 ('Two-Step Early Adult Urban') are characterised by similar timing of migration but different number of destinations. Migrants in Cluster 1 ('One-Step Early Adult Urban'), Cluster 3 ('Adolescent Urban'), and Cluster 4 ('Middle Adult Urban') are differentiated by the timing of migration but share the same frequency—one urban destination. Clusters 5 and 6 ('Rural' and 'Return') are marked by their brief urban stints, followed by rural migration or return, respectively. Cluster 7 ('Transient Urban') is characterised by the greatest frequency of migration. Note that by design, the state distribution plots do not include missing as a state. The timing of first migration for each migration pattern is further illustrated by kernel density plots in Figure A1.

3.4 | Methods and variables for studying determinants of migration patterns

Results from our sequence analysis provide the foundation for substantive analysis regarding the determinants and consequences of migration. We use multinomial logistic regression to estimate membership in each migration cluster as a function of individual and hometown characteristics measured at or before the start of migration. Rural stayers (nonmigrants) are the base category. This analysis sheds

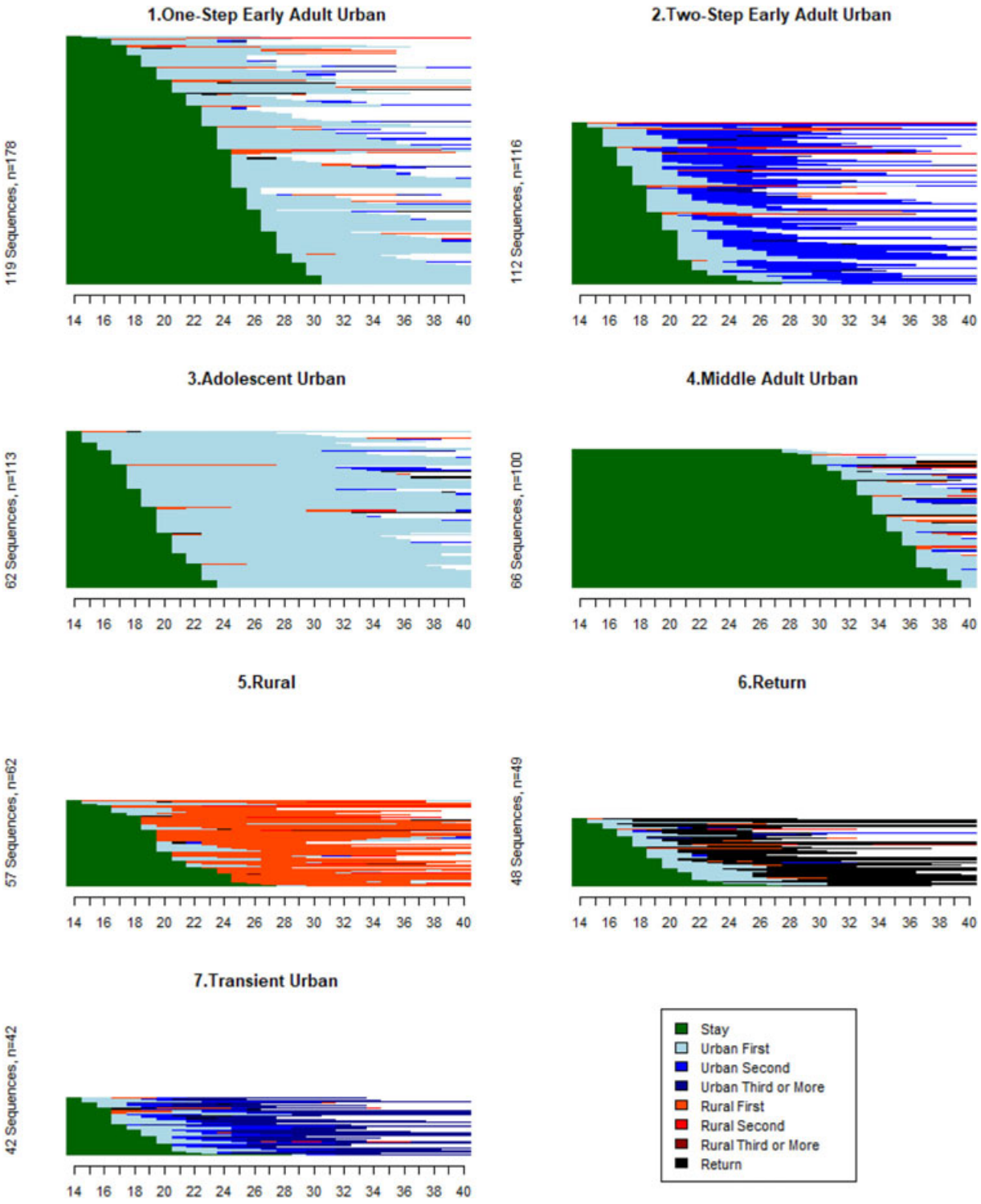

FIGURE 1 Sequence index plots of migration trajectory clusters, IMHC 2008 (weighted frequencies are shown)

light on factors that determine the migration trajectories. Individual-level factors include the respondent's birth year, gender, education, father's education, number of books at age 14, father's reading behaviour when the respondent was age 14, and protein intake at age 14. Birth year helps adjust for cohort differences in migration patterns (A. Bernard et al., 2019; Z. Liang, 2004; Shen, 2012). Gender plays a significant role in migration processes (C.C. Fan, 2000, 2004) and migrants' economic attainment (Huang, 2001; X. Meng, 1998). Human capital also affects migration patterns and economic outcomes (Hu et al., 2011; H. Li & Zahniser, 2002; Z. Liang, 2004; X. Wu & Treiman, 2007; X. Yang & Guo, 1999). To harmonise the respondent's own education and father's education, we categorise education into

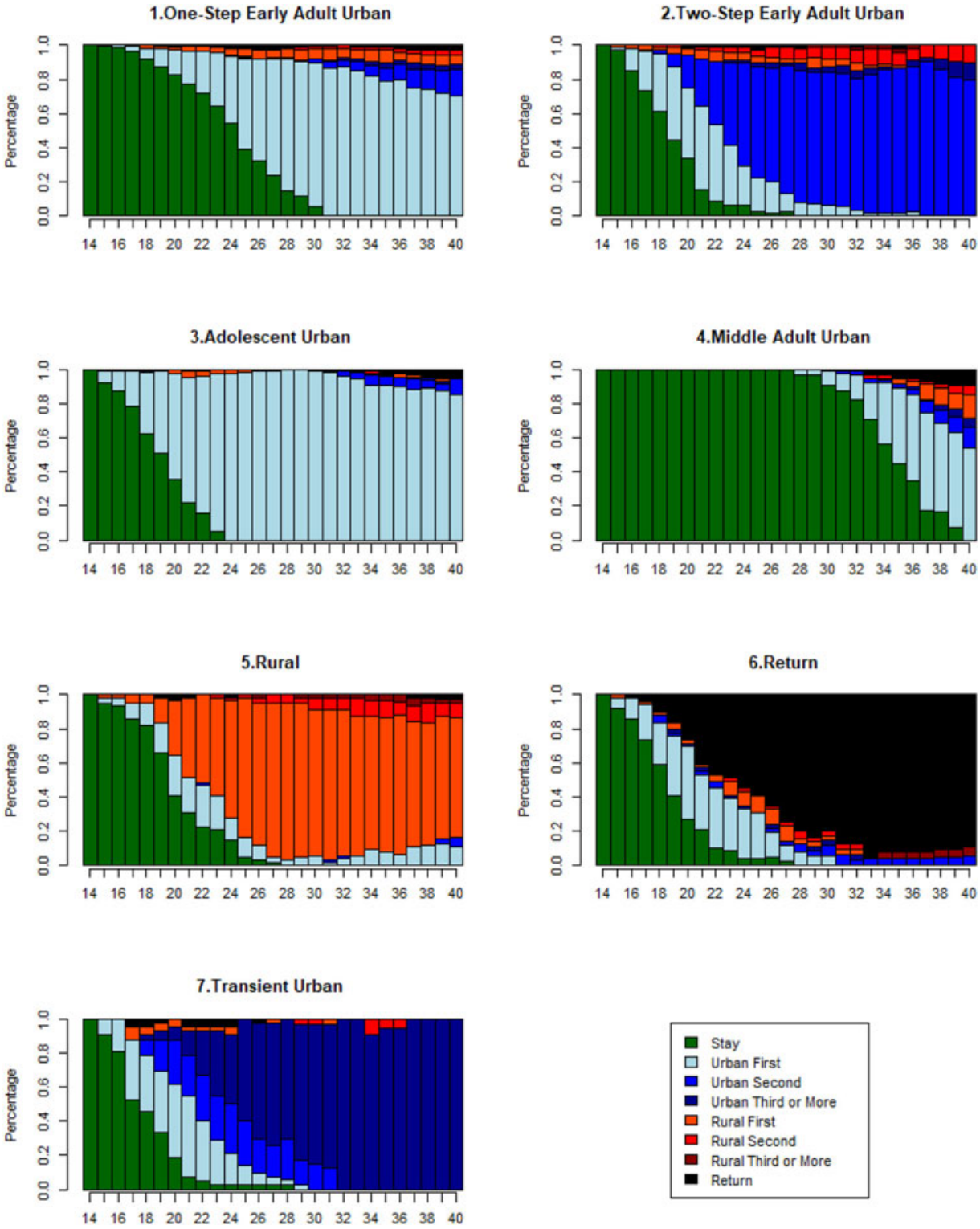

FIGURE 2 State distribution plots by migration trajectory clusters, IMHC 2008

four groups: (i) no school; (ii) less than primary school; (iii) primary school but less than middle school; and (iv) middle school or higher. The number of books at home can significantly shape one's socio-economic success by conferring competencies, skills, and knowledge (M.D. Evans et al., 2010; M.D.R. Evans et al., 2015). It is a categorical variable with four categories: '0 books', '1–9 books', '10–19 books', and '20 books or more'. We also control for whether the respondent's father ever read a book. Protein intake at age 14 is a proxy for family economic background and health (D.J. Treiman, 2012); the latter can be an important determinant of migration (Y. Lu & Qin, 2014). This

variable equals 1 if the respondent consumes protein (i.e., meat, fish, or milk) once or more a week.

We also control for hometown characteristics, including the distance between the hometown and the county seat and school availability, which is measured by whether the hometown has one, two, or all three school types (primary, middle, and high school). The distance variable taps into access to information and transportation, which could facilitate rural-to-urban migration. It is constructed using the question: 'When you moved away from here, how long would it take for a round trip to the county seat?' The response categories are 'need a day or more to return', 'need a half day or more to return', or 'lived in the county seat or larger city'. The number of school types is a good indicator of public provision in the hometown, which potentially affects migration (S. Fan & Zhang, 2004; Luo et al., 2007).

3.5 | Methods and variables for studying the consequences of migration patterns

Lastly, we investigate how migration trajectories in early to midlife shape migrants' occupational attainment using occupational status as the outcome and the trajectory clusters as the main explanatory variable. We address potential endogeneity in this analysis because individuals may self-select into different migration trajectories (based on abilities, motivation, etc.) that may also affect occupational attainment. We correct for potential selection bias using multinomial selection models that follow the counterfactual framework and estimate the average treatment effect on the treated (ATT; Adams & Cuecuecha, 2013; Parvathi & Waibel, 2016). With β_i as the treatment effect for individual i and D_i as the treatment indicator (i.e., equal to 1 if treated and 0 otherwise), then the average treatment effect (ATE) will be defined as $E[\beta_i]$, and ATT will be defined as $E[\beta_i | D_i = 1]$. As a causal estimate, ATT provides the average treatment effect for the subpopulation of treated individuals (i.e., those who experience a certain migration trajectory). Although an instrumental variable approach also may be used to estimate the causal effects of the migration patterns on occupational attainment, applying an instrumental variable approach would be extremely difficult in our case. We have eight migration patterns including stayers. This would require at least seven instruments, which would be quite hard to identify; indeed, it is not uncommon to fail to find only one suitable instrument.

We proceed with a multinomial logit selection model developed by Dubin and McFadden (1984). This approach outperforms other multinomial selection methods in efficiency (Bourguignon et al., 2007). We use a more recent augmented version of this method that relaxes the zero correlation assumption (between error terms in the two stages) in the original model and provides more robust results (Bourguignon et al., 2007). The detailed estimation procedures are discussed in Appendix C. We use the distance between the hometown and the county seat and hometown school availability as the exclusion restrictions in the selection model because these characteristics are unlikely to be associated with migrants' occupational attainment independent of their impacts on migration patterns. Then, we use the coefficients obtained in the multinomial selection models to derive our counterfactual models of occupational outcomes and estimate the ATT.

In this analysis, the main outcome variable is occupational status, measured by the widely used International Socio-Economic Index (ISEI; H.B. Ganzeboom et al., 1992; H.B. Ganzeboom & Treiman, 1996). The ISEI has been shown to be a reliable measure of fine-grained occupational status across settings, including in China (Y. Lu & Treiman, 2008; X. Wu & Treiman, 2004). A strength of ISEI is that it allows for more precisely capturing occupational mobility (upward vs. downward) than aggregated occupational categories and can be analysed in a linear regression framework. In addition to studying ISEI measured at the time of the survey, we also construct a measure of life-course ISEI score mobility using occupational histories. Specifically, using a growth curve model, we derive the growth rate (slope) of an individual's occupational trajectory over the same observation window as the migration trajectory and use it as the dependent variable. This measure is a raw estimate of each migrants' occupational trajectory. It is adjusted for controls when used as the dependent variable. A higher slope indicates greater occupational mobility.[5]

We control for all the variables included in the analysis of the determinants of migration trajectories except for the two variables about hometown characteristics (because they are used as exclusion restrictions). Additionally, when the dependent variable is the current ISEI score, we control for current individual characteristics measured at the time of the survey. These variables include household size, health status, and *hukou* status. We do not control for marital status as 94% of the respondents are married. Household size is a continuous variable that measures the total number of co-resident family members. Health status is measured as a categorical variable with four categories: 'poor', 'fair', 'good', and 'excellent'. It can affect migrants' labour force participation and occupational outcomes (Qin et al., 2015). *Hukou* status is a critical factor shaping the occupational attainment of rural-to-urban migrants (Z. Liang, 2004; X. Wu & Treiman, 2007). We measure both whether the migrant holds a non-agricultural *hukou* and whether he or she holds a local *hukou* in the place of destination. Moreover, we control for the size of the current destination (the hometown for return migrants), which could determine migrants' occupational opportunities (K.H. Zhang & Song, 2003; Zhu, 2002). We use the 10-category classification identified earlier to capture heterogeneity in size.

4 | RESULTS

4.1 | Descriptive statistics

Table 1 provides comparisons across the seven clusters by timing, direction, number of migratory trips, distance, and the reason for migration.[6] First, rural migration is highly driven by marriage-related reasons. About 45% of migrants in Cluster 5 moved for marriage. This is consistent with patrilocal residential patterns characteristic of rural China until recently, in which the bride moves to the husband's village

TABLE 1 Characteristics of different migration patterns

Characteristics	Migration patterns						
	One-step early adult urban	Two-step early adult urban	Early adolescent urban	Middle adult urban	Rural	Return	Transient urban
Direction	Urban	Urban	Urban	Urban	Rural	Return	Urban
Typical (mode) number of migratory trips	1	2	1	1	2	2	3+
Average age (first trip)	23.6	18.8	18.9	33.4	20.5	18.7	17.8
Distance (first trip)	37% cross[a]	37% cross[a]	28% cross[a]	43% cross[a]	29% cross[a]	40% cross[a]	50% cross[a]
Reason (first trip[b])	Work (52%) Marriage (25%) Family (14%)	Work (56%) Educ/train (28%)	Work (53%) Marriage (30%) Family (10%)	Work (65%) Marriage (12%) Family (10%)	Marriage (45%) Work (39%)	Work (65%) Military (25%)	Work (62%) Educ/train (17%) Military (17%)
Reason (main)[c]	-[d]	Work (71%) Marriage (14%) Family (10%)	-[d]	-[d]	-[d]	Family (47%) Work (45%)	Work (67%) Marriage (10%)
N	178	116	113	100	62	49	42

[a]'Cross' refers to cross-province.
[b]The percentages sometimes add to less than 100% because only the numerically important reasons are shown.
[c]Main migration refers to the trip with the longest duration. Among all migration patterns with more than one migration episode, there is often one migration episode that tends to be notably longer than the other migration episode(s). For example, for 'Two-Step Early Adult Urban', we find that the second migration episode tends to be the main one. For 'Rural' and 'Return' migration, we find that the rural or return migration tends to be the main episode. For 'Transient Urban', we observe that the third migration episode tends to be the main one.
[d]There is no entry because 'first' and 'main' reason are the same for these groups of migrants.

and, often, into his extended family household (C.C. Fan & Huang, 1998; C.C. Fan & Li, 2002). Second, all types of rural-to-urban migration (including return migration) is mainly labour migration; the share of labour migration ranges from 52% to 71%. Third, Clusters 1 and 3 include a substantial amount of marriage migration, presumably migration from a village to an urban area to join a spouse. Fourth, 28% of two-step early adult urban migrants (Cluster 2) and 17% of transient urban migrants (Cluster 7) first migrated for education/training purposes. Their later trips were primarily job related. This suggests that educated migrants are more likely to explore multiple urban destinations. Finally, among different types of rural-to-urban migration, adolescent urban migration (Cluster 3) involves shorter distances (fewer cross-province trips). Not surprisingly, adolescent migrants tend to move less far away from their hometown than do those who migrate at older ages.

Table 2 shows summary statistics for variables used in the regression analyses by migration trajectories. There are two general observations: (i) the characteristics of migrants differ substantially across migration trajectories, and (ii) although all clusters of migrants on average have higher current occupational status than rural stayers (those who never migrated), the gain in occupational status varies by trajectory. Also, the higher growth (steeper slopes) relative to stayers is true of most of the migrant clusters but not of those who migrated to urban areas as adolescents or who migrated to other rural areas. Note that the ISEI slope is a within-individual slope. Therefore, even if the current ISEI is significantly higher, the overall growth in occupational trajectory may not necessarily be so. These descriptive statistics should be interpreted with caution because other potential confounding factors are not controlled.

4.2 | Determinants of migration patterns

We next examine the determinants of migration trajectories in a regression framework. Table 3 presents multinomial regression results. First, demographic factors matter. Rural-to-urban migration exponentially increased, beginning in the early 1980s and continuing beyond the date of the survey (K.W. Chan, 2013; Zheng & Yang, 2016). What we show is that this increase drew mainly from those who migrated as young adults (Clusters 1 and 2), including those who already had returned (Cluster 6). Also, women were much more likely than men to be in clusters (1, 3, and 5) that were driven by marriage migration (see Table 1) and to migrate early. Women's propensity for early migration perhaps reflects family strategies promoting female labour migration at young ages to support the education of their male siblings (C.C. Fan & Huang, 1998; Gruijters & Ermisch, 2019).

Among socio-economic factors, education does not have a clear impact. This is consistent with previous research that shows either a

TABLE 2 Summary statistics

Variables	Mean (SD) or proportion								
	Total	Stay	One-step early adult urban	Two-step early adult urban	Early adolescent urban	Middle adult urban	Rural	Return	Transient urban
Socio-economic outcome									
ISEI score	35.0 (14.1)	30.9 (12.9)	40.6 (14.8)	42.0 (14.2)	36.4 (12.8)	35.6 (13.7)	31.2 (12.9)	33.1 (12.8)	39.8 (13.5)
ISEI slope	0.15 (0.92)	0.09 (0.54)	0.25 (1.32)	0.29 (1.33)	0.08 (0.69)	0.19 (0.78)	0.11 (0.79)	0.16 (1.54)	0.20 (0.90)
Socio-economic background									
Birth year	1964 (11.11)	1959 (10.15)	1969 (10.2)	1971 (10.4)	1962 (10.0)	1962 (8.50)	1962 (10.53)	1969 (11.65)	1968 (12.1)
Female	0.50	0.42	0.65	0.50	0.61	0.44	0.74	0.31	0.33
Education									
No school	0.07	0.09	0.06	0.01	0.07	0.05	0.08	0.02	0.02
Less than primary	0.13	0.17	0.09	0.05	0.07	0.19	0.19	0.14	0.05
Primary	0.23	0.25	0.17	0.14	0.26	0.25	0.37	0.29	0.12
Middle or more	0.57	0.49	0.67	0.80	0.60	0.51	0.36	0.55	0.81
Father's education									
No school	0.35	0.45	0.20	0.16	0.31	0.42	0.45	0.35	0.26
Less than primary	0.27	0.28	0.29	0.24	0.29	0.27	0.24	0.25	0.14
Primary	0.21	0.15	0.26	0.34	0.23	0.24	0.19	0.14	0.38
Middle or more	0.17	0.13	0.25	0.27	0.17	0.07	0.11	0.27	0.21
Number of books at age 14									
0 book	0.20	0.24	0.20	0.10	0.18	0.18	0.28	0.20	0.09
1–9 books	0.30	0.36	0.27	0.23	0.29	0.21	0.29	0.22	0.17
10–19 books	0.24	0.20	0.21	0.28	0.32	0.31	0.19	0.33	0.24
20 + books	0.26	0.20	0.32	0.39	0.21	0.30	0.24	0.25	0.50
Father's book Reading									
Does read	0.29	0.24	0.35	0.37	0.38	0.20	0.18	0.27	0.41
Protein intake at age 14	0.32	0.20	0.44	0.51	0.30	0.35	0.29	0.35	0.55
Home hierarchy									
Village	0.75	0.79	0.68	0.71	0.65	0.72	0.97	0.94	0.57
Home distance to the county seat									
More than a half day	0.14	0.11	0.15	0.19	0.16	0.17	0.10	0.14	0.17
Half a day	0.68	0.69	0.66	0.63	0.61	0.71	0.87	0.76	0.67
In county seat	0.18	0.20	0.19	0.18	0.23	0.12	0.03	0.10	0.17

(Continues)

TABLE 2 (Continued)

Variables	Mean (SD) or proportion								
	Total	Stay	One-step early adult urban	Two-step early adult urban	Early adolescent urban	Middle adult urban	Rural	Return	Transient urban
Socio-economic status									
Household size	4.02 (2.10)	4.02 (1.71)	3.92 (2.36)	3.88 (2.65)	3.66 (1.70)	3.88 (2.38)	4.42 (1.92)	4.94 (2.85)	4.48 (2.37)
Health status									
Poor	0.13	0.14	0.14	0.06	0.13	0.12	0.19	0.12	0.14
Fair	0.43	0.44	0.40	0.41	0.49	0.44	0.44	0.39	0.36
Good	0.28	0.27	0.26	0.41	0.26	0.27	0.26	0.29	0.26
Excellent	0.16	0.15	0.20	0.12	0.12	0.17	0.11	0.20	0.24
Local Hukou	0.77	0.99	0.56	0.59	0.70	0.53	0.79	0.86	0.60
Non-agricultural Hukou	0.31	0.19	0.40	0.45	0.56	0.37	0.21	0.12	0.45
Destination hierarchy									
Small Village	0.25	0.46	0.06	0.06	0.01	0.12	0.39	0.33	0.02
Ordinary village	0.15	0.26	0.03	0.03	0.02	0.06	0.21	0.37	0.05
Large village	0.05	0.07	0.03	0.03	0.00	0.03	0.08	0.06	0.02
Xiang township seat	0.01	0.02	0.01	0.01	0.01	0.00	0.03	0.02	0.00
Zhen township seat	0.15	0.09	0.20	0.17	0.27	0.19	0.06	0.06	0.17
County seat	0.03	0.01	0.03	0.04	0.01	0.08	0.02	0.02	0.10
County-level City	0.03	0.00	0.08	0.05	0.05	0.04	0.06	0.00	0.05
Prefecture-level City	0.15	0.04	0.25	0.34	0.28	0.22	0.05	0.02	0.24
Province capital	0.11	0.03	0.15	0.19	0.26	0.11	0.06	0.10	0.19
Provincial-level City	0.07	0.01	0.16	0.09	0.09	0.15	0.03	0.02	0.17
N	1,112	452	178	116	113	100	62	49	42

Note: See text for details on variable definitions.

TABLE 3 Multinomial logistic regression of determinants of migration patterns

	Patterns						
Variables	**One-step early adult urban**	**Two-step early adult urban**	**Early adolescent urban**	**Middle adult urban**	**Rural**	**Return**	**Transient urban**
Birth year	0.070*** (0.011)	0.078*** (0.014)	0.003 (0.012)	0.008 (0.012)	0.016 (0.016)	0.086*** (0.019)	0.038* (0.020)
Female	0.826*** (0.201)	0.277 (0.232)	0.792*** (0.226)	0.131 (0.239)	1.375*** (0.322)	−0.633^{+} (0.340)	−0.461 (0.367)
Education (ref = no school)							
Less than primary	−0.330 (0.469)	0.853 (1.117)	−0.457 (0.548)	0.792 (0.553)	0.610 (0.585)	1.024 (1.106)	−0.217 (1.268)
Primary	−0.443 (0.439)	0.855 (1.073)	0.307 (0.467)	0.390 (0.551)	0.841 (0.560)	0.899 (1.087)	−0.254 (1.166)
Middle or more	0.106 (0.420)	1.693 (1.053)	0.493 (0.459)	0.395 (0.545)	0.286 (0.583)	0.758 (1.082)	0.546 (1.106)
Book at 14 (ref = 0)							
1–9	−0.314 (0.283)	0.141 (0.407)	−0.036 (0.326)	−0.136 (0.356)	−0.282 (0.382)	−0.468 (0.477)	−0.096 (0.675)
10–19	−0.296 (0.315)	0.525 (0.414)	0.426 (0.344)	0.892* (0.359)	−0.060 (0.445)	0.289 (0.478)	0.543 (0.660)
20+	−0.233 (0.316)	0.430 (0.417)	−0.075 (0.382)	0.961* (0.376)	0.380 (0.450)	−0.292 (0.523)	1.054^{+} (0.639)
Father's education (ref = no school)							
Less than primary	0.496^{+} (0.274)	0.343 (0.352)	0.186 (0.294)	0.037 (0.298)	−0.206 (0.369)	−0.215 (0.428)	−0.653 (0.561)
Primary	0.715* (0.304)	0.923** (0.358)	0.406 (0.332)	0.331 (0.334)	−0.045 (0.426)	−0.372 (0.518)	0.688 (0.483)
Middle or more	0.491 (0.343)	0.461 (0.405)	0.078 (0.394)	−0.836^{+} (0.499)	−0.351 (0.540)	0.171 (0.508)	−0.310 (0.596)
Father reads (ref = does not read)	−0.133 (0.237)	−0.241 (0.269)	0.302 (0.269)	−0.492 (0.315)	−0.440 (0.408)	−0.265 (0.406)	0.148 (0.405)
Protein intake at age 14	0.533* (0.213)	0.674** (0.247)	0.336 (0.255)	0.709** (0.259)	0.383 (0.330)	0.207 (0.354)	1.052** (0.369)
Home distance to the county seat (ref = half a day)							
More than a half day	0.391 (0.286)	0.761* (0.316)	0.484 (0.316)	0.399 (0.321)	−0.450 (0.471)	0.191 (0.457)	0.682 (0.475)
In county seat	−0.168 (0.252)	−0.293 (0.301)	0.114 (0.271)	−0.595^{+} (0.344)	−2.176** (0.739)	−0.869^{+} (0.507)	−0.523 (0.461)
Number of school types (ref = primary only)							
Primary and middle	−0.161 (0.307)	0.327 (0.373)	−0.236 (0.320)	−0.059 (0.329)	−0.747^{+} (0.453)	0.294 (0.462)	2.329* (1.071)
Primary, middle, and high	0.262 (0.255)	0.509 (0.323)	−0.189 (0.270)	−0.057 (0.288)	0.083 (0.341)	0.109 (0.426)	2.607* (1.034)
Constant	−138.950*** (22.257)	−157.510*** (27.505)	−8.203 (23.245)	−17.355 (24.068)	−33.900 (30.453)	−170.929*** (36.470)	−81.008* (38.171)
Observations	1,112	1,112	1,112	1,112	1,112	1,112	1,112

Note: Coefficients are logits (log-odds); stayers—non-migrants—is the reference category. Standard errors in parentheses.
***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. $^{+}p < 0.1$.

null or a complicated curvilinear relationship between education and migration (H. Li & Zahniser, 2002; X. Yang & Guo, 1999). By contrast, the number of books at age 14 seems to be positively associated with middle adulthood urban migration (Cluster 4). This group of migrants may be able to find good local employment opportunities (teachers and village officials) early on in life and seek even better jobs in cities later on after accumulating human and financial capital. Also, father's education is related to early adult rural-to-urban migration (Clusters 1 and 2). Fathers with more education (middle school or above) may encourage their children to pursue education before migration. Moreover, individuals with more frequent protein intake at age 14 are more likely to migrate as adults (Clusters 1, 2, and 4) and to migrate more often (Cluster 7). This result is consistent with the healthy migrant hypothesis that healthier people are more likely to migrate (Y. Lu & Qin, 2014).

Beyond individual characteristics, hometown features affect migration patterns. Interestingly, living more than half a day from the county seat promotes multiple-trip migration (Clusters 2 and possibly Cluster 7, although the coefficient for Cluster 7 is not statistically significant, probably due to the small size of the cluster); this is perhaps because the cost of returning home is so great. Not surprisingly, those living in county seats are particularly unlikely to migrate to rural areas—as others have observed (Hu et al., 2011; Zhao, 2003), the general pattern in China is to migrate to larger places. Interestingly, communities with more types of schools tend to produce 'transient' migrants (Cluster 7)—those who have moved to three or more places. The reason for this is not clear.

4.3 | Consequences of migration patterns

Tables 4 and 5 show how various early-life factors affect migrants' occupational status (ISEI) and occupational trajectories (ISEI trajectory slope) using a multinomial logit selection model (Bourguignon et al., 2007; Dubin & McFadden, 1984). The most important variables in both tables are the selection terms—lambdas. Lambda *j* represents the selection correction variables related to migration pattern *j* (i.e., error correlations between the two stages of estimation). The results show that many of the lambdas are significantly different from zero, indicating that the estimates would be biassed if selection is not considered. Different from the Heckman selection model, in the multinomial logit selection model, lambdas are seldom interpreted independently. This is because the interpretation is not clear given that we have various error correlations between the two stages of estimation. To suggest meaningful interpretations, one needs to assume the latent factors captured in each error correlation. As a result, studies that have applied this selection model have only focused on the significances of the lambdas and not their size or direction.

To estimate the association between migration patterns and occupational outcomes while accounting for potential selection bias, we proceed under the counterfactual framework. Table 6 shows the average treatment effects on the treated (ATT) estimated from the difference between the actual outcome and the counterfactual outcome. With eight migration patterns (including stayers), we construct seven counterfactuals $E(ISEI_{istay} | Pattern_i = j)$, which represents what the ISEI and ISEI trajectory would be for migrants who experienced migration pattern *j* had they stayed at their rural homes. It is calculated using the outcome equations for migrants in migration pattern *j* relative to rural stayers, using coefficients obtained from the multinomial selection models in Tables 4 and 5. This counterfactual strategy thus accounts for potential selection bias due to unobserved heterogeneity. The estimated outcome is $E(ISEI_{ij} | Pattern_i = j)$, which represents the ISEI for migrants who experienced migration pattern *j* had they chosen pattern *j* (i.e., actual case). Then, our ATT is calculated as follows:

$$ATT(j) = E(ISEI_{ij} | Pattern_i = j) - E(ISEI_{istay} | Pattern_i = j).$$

Table 6 shows the estimated outcomes, counterfactual outcomes, and ATT for migrants in different clusters. The results are generally consistent for the ISEI and ISEI trajectory. Several results are worth highlighting. First, migrants in the Rural and Return migration clusters (5 and 6) did not attain higher occupational attainment or experience greater occupational mobility than had they stayed home. This could reflect the limited non-agricultural opportunities in rural areas. Previous research suggests that return migrants may undertake entrepreneurship upon returning (S. Démurger & Xu, 2011; Ma, 2002), which would lead to a higher ISEI than resuming agriculture. Our finding that return migrants are indistinguishable from rural nonmigrants suggests that return migrant entrepreneurship is likely the exception rather than the rule, especially for migrants who return by age 40. Of course, it may be the case that return migrants had higher status jobs when they were out for work, that is, during their period as migrants, or that regardless of the status of the jobs they held they earned more than had they stayed home. This must be true on average because, otherwise, the motivation to migrate would have dissipated in rural communities as information about the lack of urban opportunities spread.

Second, non-transient adult urban migration (Clusters 1, 2, and 4) is associated with better occupational outcomes, measured by both occupational status and occupational trajectories. Migrants in Clusters 1, 2, and 4 had experienced a growth rate twice as high as if they did not migrate. For example, migrants who experienced one-step early adult migration (Cluster 1) had a slope of 0.248, with a counterfactual slope of only 0.117 had they not migrated. Note that the average occupational outcome varies by the specific migration pattern. On the one hand, one-step early adult urban migration (Cluster 1) is the most beneficial migration pattern as measured by occupational status, which may reflect high job stability and a resulting greater adaptation to the destination. On the other hand, two-step early adult urban migration (Cluster 2) is particularly conducive to occupational mobility. This suggests that accumulation of work experience and human capital can lead to greater job opportunities in a new destination.

Third, adolescent or transient urban migration (Cluster 3 and 7) may not be occupationally beneficial. As discussed above, adolescent urban migration (Cluster 3) is disproportionately composed of within-province migration that may not significantly boost employment

TABLE 4 ISEI estimates (selection corrected) for all migration patterns, using DMF method

	Patterns							
VARIABLES	Stay	One-step early adult urban	Two-step early adult urban	Early adolescent urban	Middle adult urban	Rural	Return	Transient urban
Birth year	−0.596* (0.280)	0.310 (0.823)	−0.355 (1.244)	0.845 (0.722)	−0.174 (0.889)	1.804 (10.038)	2.447 (8.658)	2.912 (2.751)
Female	−4.081 (4.826)	−3.590 (9.050)	3.251 (14.898)	0.715 (14.600)	−0.052 (15.078)	18.491 (102.792)	3.657 (208.872)	72.351 (58.498)
Education (ref = no school)								
Less than primary	6.604 (4.382)	1.817 (11.210)	−29.805 (26.402)	−6.434 (20.149)	−8.552 (31.552)	−15.114 (146.566)	46.698 (203.205)	−61.597 (62.153)
Primary	3.619 (4.353)	6.914 (11.985)	−16.203 (23.983)	−1.844 (13.962)	16.096 (23.162)	8.345 (135.111)	16.782 (145.019)	32.704 (81.072)
Middle or more	6.569 (4.577)	8.923 (10.144)	−12.305 (25.852)	8.330 (14.251)	16.713 (24.968)	11.283 (146.403)	29.061 (168.823)	19.268 (77.497)
Book at 14 (ref = 0)								
1–9	2.759 (3.063)	5.889 (6.231)	1.848 (13.969)	0.512 (9.222)	2.687 (12.821)	−5.984 (47.094)	−35.767 (209.488)	−28.310 (67.333)
10–19	8.279+ (4.887)	7.160 (10.581)	7.169 (16.130)	3.972 (13.465)	−9.382 (16.352)	−4.101 (70.344)	−8.457 (193.469)	−38.262 (56.635)
20+	6.560 (4.873)	−2.587 (9.624)	5.297 (15.078)	−6.011 (14.597)	−20.925 (22.466)	−20.819 (75.873)	−21.668 (274.364)	−85.951 (70.264)
Father's education (ref = no school)								
Less than primary	−2.489 (2.649)	1.159 (8.945)	3.528 (13.976)	−0.262 (9.701)	−4.458 (11.383)	20.821 (46.997)	9.372 (80.186)	32.348 (79.243)
Primary	−0.125 (4.058)	−6.423 (8.482)	8.816 (16.202)	6.283 (11.310)	−5.367 (15.559)	11.705 (54.252)	−6.699 (130.706)	4.547 (67.578)
Middle or more	−6.057 (5.019)	0.835 (11.315)	1.173 (16.754)	9.789 (14.104)	15.323 (22.847)	26.575 (65.032)	12.109 (172.848)	45.527 (90.783)
Father reads (ref = does not read)	0.405 (3.562)	4.279 (7.028)	5.128 (9.965)	5.055 (11.444)	9.913 (12.587)	2.005 (132.484)	−9.301 (261.143)	62.043+ (33.249)
Protein intake at age 14	2.636 (3.725)	3.567 (6.118)	−2.742 (10.820)	−3.844 (8.744)	−13.733 (10.229)	−2.061 (65.889)	9.243 (84.076)	−15.453 (59.831)
Household size	−0.496 (0.346)	−0.224 (0.521)	−0.978 (0.803)	−1.137 (0.986)	−0.495 (0.793)	−0.811 (7.097)	0.132 (16.910)	3.318 (7.015)
Health status (ref = poor)								
Fair	1.851 (1.358)	−1.234 (3.322)	−0.644 (9.075)	−1.170 (4.573)	10.579* (5.305)	−0.487 (30.595)	−7.802 (105.767)	29.301 (64.297)
Good	4.401** (1.530)	−5.285 (3.853)	−2.921 (8.254)	0.026 (5.481)	8.689+ (5.279)	5.586 (42.174)	5.620 (64.539)	9.380 (48.204)
Excellent	3.439+ (1.917)	−5.186 (4.361)	−9.201 (9.660)	−3.500 (8.196)	10.546 (6.775)	2.169 (52.738)	−15.685 (131.832)	13.847 (52.402)
Local Hukou	−4.085 (5.149)	−0.762 (3.117)	4.526 (3.693)	−5.388 (5.085)	−2.806 (5.037)	−6.550 (155.215)	−13.274 (130.907)	22.786 (37.719)
Non-agricultural Hukou	6.911** (2.497)	11.613*** (3.257)	4.562 (4.119)	2.536 (5.627)	12.301* (4.941)	22.770 (37.138)	−0.156 (76.596)	21.697 (32.611)
Destination hierarchy (ref = small village)								
Ordinary village	4.278** (1.521)	6.178 (11.792)	10.327 (14.118)	−14.412 (13.491)	2.976 (7.084)	−3.600 (59.155)	−4.532 (70.253)	44.796 (72.964)
Large village	3.729 (2.507)	8.085 (8.382)	18.293 (12.284)	0.000 (0.000)	−4.783 (9.534)	16.855 (33.597)	−6.417 (101.544)	−51.718 (36.460)
	6.689 (4.154)	11.872 (9.365)	12.419 (8.495)	17.796 (16.137)	0.000 (0.000)	−9.681 (35.350)	32.309 (169.722)	0.000 (0.000)

(Continues)

TABLE 4 (Continued)

VARIABLES	Patterns							
	Stay	One-step early adult urban	Two-step early adult urban	Early adolescent urban	Middle adult urban	Rural	Return	Transient urban
Township seat of a Xiang								
Township seat of a Zhen	3.746 (2.833)	12.073* (5.677)	19.732** (7.232)	0.471 (14.153)	−0.059 (7.285)	6.502 (48.541)	2.062 (72.498)	9.330 (61.835)
County seat	−6.438 (8.314)	1.101 (7.113)	12.957 (12.766)	−15.634 (14.403)	0.009 (11.251)	−7.930 (211.648)	−22.444 (76.248)	−6.081 (74.395)
County-level City	−1.770 (2.421)	7.114 (5.938)	6.378 (8.661)	−7.521 (14.841)	5.590 (11.677)	−7.271 (213.900)	0.000 (0.000)	13.125 (75.158)
Prefecture-level City	0.465 (3.979)	8.137 (5.862)	15.797* (6.645)	1.984 (14.888)	−6.692 (6.841)	4.658 (88.115)	−32.161 (350.469)	5.555 (65.849)
Province capital	4.850 (5.172)	10.592 (6.714)	15.205* (7.381)	−0.587 (15.119)	5.592 (6.968)	2.973 (52.120)	1.301 (89.713)	4.211 (61.937)
Provincial-level City	−4.244 (5.500)	10.289 (7.521)	14.843 (11.795)	3.261 (16.788)	−7.634 (7.929)	−8.416 (144.600)	−24.733 (116.414)	37.468 (75.521)
Selection terms								
Lambda 1	0.085 (6.072)	−6.466 (39.082)	−3.808 (56.628)	7.037 (59.023)	192.753** (65.133)	2.811 (401.563)	−199.090 (422.343)	−441.668*** (109.097)
Lambda 2	−34.759 (22.877)	−8.800 (9.947)	−38.000 (68.950)	1.008 (62.041)	27.903 (67.368)	106.659 (497.909)	139.443 (412.514)	−18.360 (119.926)
Lambda 3	−7.590 (22.547)	25.933 (37.783)	3.143 (14.174)	86.051^{+} (49.736)	138.77^{+} (73.202)	90.367 (529.516)	−212.959 (797.499)	−296.625*** (77.196)
Lambda 4	−1.450 (23.732)	9.934 (47.120)	53.093 (66.827)	6.331 (14.119)	115.497 (79.393)	153.215 (764.887)	−96.451 (965.729)	365.988*** (77.324)
Lambda 5	40.530* (18.485)	−24.691 (40.087)	1.358 (58.821)	−66.656 (60.560)	−6.074 (14.534)	−32.293 (218.639)	80.128 (1,060.038)	−577.840*** (165.566)
Lambda 6	−16.207 (13.062)	10.887 (26.621)	3.048 (41.767)	29.787 (27.256)	122.644** (41.966)	6.647 (61.723)	−140.933 (1,104.772)	84.758 (89.333)
Lambda 7	−26.794 (20.924)	26.134 (31.191)	−40.302 (56.887)	52.555 (64.481)	80.020 (63.466)	100.944 (440.144)	16.045 (108.936)	281.097* (110.051)
Lambda 8	−12.458 (17.105)	−30.032 (28.463)	−10.096 (50.211)	−19.134 (47.165)	42.478 (47.459)	−45.330 (318.355)	−3.566 (319.265)	−8.917 (49.055)
Constant	1,183.966* (541.760)	−569.570 (1,658.100)	713.748 (2,501.797)	−1,618.331 (1,400.153)	604.029 (1,699.548)	−3,450.642 (19,743.711)	−4,970.885 (17,340.345)	−6,080.802 (5,441.094)
Observations	1,112	1,112	1,112	1,112	1,112	1,112	1,112	1,112

Note: Standard errors in parentheses.
***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. $^{+}p < 0.1$.

TABLE 5 ISEI trajectory estimates (selection corrected) for all migration patterns, using DMF method

	Patterns							
Variables	Stay	One-step early adult urban	Two-step early adult urban	Early adolescent urban	Middle adult urban	Rural	Return	Transient urban
Birth year	0.005 (0.012)	−0.093 (0.090)	−0.013 (0.091)	0.043 (0.040)	−0.009 (0.045)	0.082 (0.101)	−0.001 (0.435)	−0.026 (0.495)
Female	−0.065 (0.183)	−0.171 (1.200)	−0.752 (1.161)	−0.516 (0.704)	−0.799 (0.835)	−0.203 (1.435)	1.257 (5.489)	2.301 (12.721)
Education (ref = no school)								
Less than primary	−0.006 (0.193)	0.586 (1.241)	1.477 (1.756)	0.202 (0.653)	−0.065 (1.529)	0.897 (1.933)	−0.839 (5.925)	−1.879 (10.466)
Primary	0.037 (0.185)	1.711 (1.229)	1.616 (1.756)	0.058 (0.545)	−1.116 (1.386)	−0.060 (1.795)	−3.260 (4.012)	−1.958 (9.448)
Middle or more	0.061 (0.218)	1.651 (1.190)	1.718 (1.731)	0.261 (0.532)	−1.379 (1.491)	0.220 (1.613)	−3.632 (4.664)	−2.540 (11.721)
Book at 14 (ref = 0)								
1–9	−0.053 (0.117)	0.624 (0.773)	0.692 (1.217)	0.029 (0.400)	−0.184 (0.596)	−0.739 (0.884)	1.541 (4.275)	−0.451 (8.728)
10–19	−0.002 (0.186)	1.286 (1.079)	0.160 (1.179)	−0.149 (0.623)	0.554 (0.965)	−0.196 (1.563)	3.396 (4.131)	−0.391 (5.873)
20+	−0.199 (0.221)	0.656 (1.090)	0.436 (1.223)	0.039 (0.632)	0.638 (1.056)	0.215 (1.336)	2.394 (7.352)	−0.906 (6.343)
Father's education (ref = no school)								
Less than primary	0.090 (0.127)	−0.269 (0.727)	−0.683 (1.004)	0.013 (0.418)	0.207 (0.548)	0.529 (1.046)	1.697 (3.112)	0.690 (15.125)
Primary	0.098 (0.148)	0.014 (0.817)	−0.327 (0.980)	0.124 (0.582)	0.254 (0.712)	0.598 (1.118)	1.267 (4.417)	−1.008 (7.300)
Middle or more	0.203 (0.190)	0.301 (0.920)	−0.813 (1.350)	0.031 (0.589)	−0.623 (1.373)	0.822 (1.447)	0.409 (6.687)	−0.543 (10.344)
Father reads (ref = does not read)	−0.135 (0.149)	0.512 (0.651)	−0.321 (0.921)	−0.133 (0.498)	−0.802 (0.672)	−0.210 (0.956)	−1.939 (4.626)	0.833 (5.527)
Protein intake at age 14	0.041 (0.145)	0.068 (0.616)	−0.222 (0.883)	0.441 (0.462)	0.137 (0.758)	0.373 (0.916)	0.642 (3.073)	−0.490 (5.182)
Selection terms								
Lambda 1	0.064 (0.268)	0.684 (4.396)	2.330 (4.300)	4.553 (3.009)	−1.297 (3.865)	2.555 (5.713)	−15.116 (18.253)	−10.833 (22.422)
Lambda 2	0.357 (1.026)	−2.224^{+} (1.186)	−2.524 (5.612)	3.999 (2.851)	1.805 (3.328)	6.418 (7.532)	−11.273 (14.758)	2.430 (25.249)
Lambda 3	0.463 (0.974)	7.639^{+} 4.070)	0.575 (1.021)	5.218* (2.572)	−5.327 (3.759)	5.596 (5.692)	−1.925 (16.756)	−14.883 (47.121)
Lambda 4	0.208 (1.005)	7.511^{+} (4.069)	−3.934 (6.685)	−0.052 (0.671)	−6.952 (4.429)	−4.668 (7.656)	−5.932 (29.814)	9.630 (35.658)
Lambda 5	0.375 (0.959)	−4.575 (4.160)	1.780 (4.872)	2.592 (2.729)	0.847 (0.778)	6.366 (5.519)	2.595 (18.942)	−5.689 (20.595)
Lambda 6	−0.184 (0.578)	4.139^{+} (2.245)	0.686 (3.702)	0.390 (1.334)	−3.223 (2.427)	−0.086 (0.564)	5.109 (21.761)	−1.526 (49.960)
Lambda 7	0.554 (0.826)	−0.686 (3.330)	−0.209 (4.951)	2.196 (2.717)	−0.620 (3.671)	4.114 (6.808)	−0.771 (4.864)	0.272 (29.603)
Lambda 8	−1.236 (0.785)	1.692 (2.872)	−0.423 (3.605)	0.839 (1.803)	0.201 (2.230)	−6.499 (5.296)	−7.818 (15.430)	−0.678 (9.688)
Constant	−8.784 (23.159)	188.959 (182.323)	23.595 (180.358)	−77.503 (77.220)	13.205 (86.534)	−156.328 (196.806)	−10.300 (872.678)	47.622 (961.599)
Observations	1,112	1,112	1,112	1,112	1,112	1,112	1,112	1,112

Note: Standard errors in parentheses. The controls differ from the analysis using current ISEI because we can only control for the variables before migration.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. $^{+}p < 0.1$.

TABLE 6 Average treatment effects on the treated (ATT) for all migration patterns

ISEI	Migration patterns		Estimated	Counterfactual	ATT	*p* value
Current	One-step early adult urban	DMF	40.61	35.44	5.17	0.000
	Two-step early adult urban	DMF	41.97	37.77	4.20	0.000
	Adolescent urban	DMF	36.35	36.19	0.17	0.812
	Middle adult urban	DMF	35.56	32.92	2.64	0.002
	Rural	DMF	31.21	30.91	0.30	0.800
	Return	DMF	33.08	33.03	0.05	0.969
	Transient urban	DMF	39.83	36.94	2.89	0.141
Trajectory	One-step early adult urban	DMF	0.248	0.117	0.131	0.001
	Two-step early adult urban	DMF	0.285	0.141	0.144	0.001
	Adolescent urban	DMF	0.075	0.086	−0.011	0.682
	Middle adult urban	DMF	0.193	0.102	0.090	0.018
	Rural	DMF	0.105	0.071	0.034	0.576
	Return	DMF	0.157	0.114	0.043	0.789
	Transient urban	DMF	0.198	0.195	0.003	0.981

opportunities. Also, this group of migrants tends to drop out of school at an early age and thus to have limited human capital with which to pursue occupational advancement. In fact, comparing adolescent urban migrants (Cluster 3) with early adult urban migrants (Cluster 1 and 2), we find that adolescent urban migrants tend to have less education, fewer books at age 14, and lower protein intake at age 14 (see Table 2). Transient urban migrants (Cluster 7) on average moved to five destinations (the median number of destinations is 4) and stayed on average 3.7 years at each place. Such a migration pattern may reflect an inability to find or to keep a stable job. This may suggest a reverse causal order—occupational histories leading to migration patterns. However, this group of migrants also have the highest level of education and more stimulating family environments when growing up (as measured by the number of books at age 14). It is more likely that their recurring migration reflects voluntary choice rather than forced choice because they are presumably most advantaged with respect to searching for a job in any urban locale. Therefore, reverse causality is unlikely to be the primary explanation for the association. Also, our result is based on current ISEI (which is not subject to reverse causality). It is likely that these migrants have sought to leverage repeated geographical mobility for upward mobility, but with little success.

5 | CONCLUSION AND DISCUSSION

The present study moves beyond the static view that has dominated previous migration research and adopts a life course perspective. We seek to capture the migration trajectory during the observation window (ages 14–40) and to account for the substantial heterogeneity in migration patterns characterised by different timing, duration, frequency, and direction among rural-to-urban migrants in China. We do so using sequence analysis, which conceptualises migration trajectories as a serial succession of multiple migration states over time. We further examine determinants of different migration trajectories and analyse how such trajectories in early to midlife shape subsequent occupational status as well as life-course occupational mobility. Overall, results highlight the importance of understanding life-course migration patterns and their relationship to social origins and subsequent socio-economic attainment.

Specifically, we identify seven distinct yet common migration trajectories based on timing, frequency, duration, and direction, including, in descending order of frequency, (i) one-step early adult rural-to-urban migration; (ii) two-step early adult rural-to-urban migration; (iii) adolescent rural-to-urban migration; (iv) middle adult rural-to-urban migration; (v) rural migration; (vi) return migration; and (vii) transient rural-to-urban migration. These patterns reveal both stability and fluidity in rural-to-urban migration in China. There is considerably more stability in rural-to-urban migration than previously thought. Return by midlife and migration to three or more destinations between early adulthood and midlife do occur but are relatively uncommon. Many migrants may return during holidays, but such visits are not counted by our definition. Genuine circular migration and serial migration do occur but are not very common.

In addition, we find that social origins at both the individual and the community level play important roles in shaping subsequent migration trajectories. Furthermore, we show that migration patterns are associated with occupational outcomes even after adjusting for self-selection into different migration trajectories. This finding holds when we look at both occupational status (measured by ISEI) at the time of the survey and occupational trajectories. The impact of rural-to-urban migration on occupational outcomes depends on the specific migration pattern. Early adult urban migration turns out to be the most advantaged trajectory for migrants' occupational attainment by

midlife, mainly because most such migration is associated with a change from agricultural to non-agricultural hukou. In comparison, migration during adolescence and frequent migration do not seem to confer occupational rewards. Although many migrants in these two migration patterns also experience a change from agricultural to non-agricultural hukou, they may lack human capital accumulation. Compared with early adult urban migrants, adolescent urban migrants have less education and fewer books at age 14 (see Table 2). Frequent migration, featured by a relatively short stay in multiple destinations, could be associated with limited and truncated capital accumulation process. When migrants do not have sufficient human capital prior to migration (e.g., adolescent urban migration) or fail to accumulate sufficient capital at each destination (e.g., transient rural-to-urban migration), their occupational gains tend to be limited. We also find that return migration by midlife is not commonly associated with entrepreneurship.

The present study has made conceptual and methodological contributions to the migration literature. It advances our conceptualization and measurement of migration as a dynamic life-course process and offers new insight into patterns and consequences of migration. When applying our conceptual and analytic framework to the case of China, we uncover substantial diversity and complexity in rural-to-urban migration experiences and how they are linked to various demographic and socio-economic factors. The timing, duration, frequency, and direction of migration appear to be important differentiating factors for rural-to-urban migration in China.

From an empirical perspective, this study illustrates the value of sequence analysis for understanding life-course migration patterns. This method allows us to identify a finite number of empirically common but substantively distinct migration trajectories. These migration trajectories can be combined with other information to better understand the causes, processes, and consequences of migration. Important heterogeneity and complexity among migrants would be missed if a snapshot approach were taken. This could obscure the multiple determinants and consequences of migration.

Several limitations warrant discussion. First, we observe migration histories up to midlife (age 40). Ideally, we would like to observe the entire migration history over a person's life. There are reasons to expect the full life-course migration trajectories to be even more diverse and dynamic. A fruitful direction for future research would be to examine longer-term migration trajectories over the life course, ideally with long-term longitudinal data to mitigate recall bias. Second, our data were collected in 2008 and thus focus on first generation migration. Second-generation migrants in China are coming of age and may exhibit different migration patterns from the earlier generation. Further research using more recent data would help capture generational differences in migration. Third, due to the limited sample size of our data set, the last three migration patterns are based on small numbers of cases. This may reduce the reliability of the results for these three migration patterns. Future study using larger samples may alleviate this potential problem. Moreover, we define migration states based mainly on the timing, frequency, and direction of migration. Migration can be further disaggregated by the reason for migration, the distance migrated, and perhaps still other factors. Considering the reason for migration would permit integrating other life-course events and transitions (e.g., changes in family structure or *hukou* status) into understanding migration patterns. One needs larger-scale data with extremely detailed biographical information to do this. We hope this study inspires future researchers to collect and analyse detailed migration histories and associated life-course events from larger samples.

ACKNOWLEDGEMENTS

The survey on which this article is based was conducted using funds from a grant to the University of California, Los Angeles, from the National Science Foundation (0551279). The article was prepared using facilities provided to the third author by the California Center for Population Research, which is supported by funding from infrastructure grant R24HD041022 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. Lu gratefully acknowledges support from the National Institute of Child Health and Development (K01HD073318).

ENDNOTES

[1] When we refer to the 'life course' perspective or to an individual's 'life course', it should be understood that we mean a sequence of migration states that unfolds as people go through life. This need not be an individual's entire life but may restricted to a particular age range, as it does in the analysis we present here.

[2] M. Yang et al. (2020) apply sequence analysis to study internal migration patterns in China. Their paper focuses on health outcomes. Our paper differs in our attention to socioeconomic outcomes of migration patterns. We also differ in how migration states are defined. The definition of migration states in M. Yang et al. (2020) is based on the reason for migration (e.g., education, family, or job related). Our measure of migration state is based on the directionality of migration (e.g., rural, urban, and return).

[3] The *hukou* (household registration) system is an institutional arrangement that categorises Chinese citizens into rural versus urban (precisely, agricultural vs. non-agricultural) based on their place of birth and parents' *hukou* status; it is difficult to convert a rural *hukou* into an urban one (X. Wu & Treiman, 2004). We do not further distinguish people with rural *hukou* at age 14 but who lived in an urban area at that age (see D.J. Treiman, 2012, and Z. Zhang & Treiman, 2013, for further discussion of this distinction). This strategy avoids reducing the sample to a problematically small size. Sensitivity analysis that restricted the sample to those with rural residence at age 14 yields very similar results.

[4] The hometown is defined as the location where the person was born or, if different, where he or she lived at age 14. Overall, the birth place and place at age 14 are the same for 96% of the respondents.

[5] The latent growth curve models and the multinomial selection models cannot be jointly estimated because the two are not compatible. Thus, they are analysed separately. Essentially, the slope is obtained from the linear regression between ISEI and age (14–40).

[6] We do not distinguish migration status by distance and reason for migration in the sequence analysis because of insufficient sample size. Also, various reasons and distances spread across all migration trajectories. This strategy allows us to focus on the main picture characterised by timing, duration, frequency, and direction.

[7] The following cases are excluded: (1) temporary visits home from school to visit the family; (2) temporary visits home from work during the Spring

Festival, even if the visit lasted more than a month; (3) traveling during the May 1, October 1, or other holidays; (4) if the respondent had moved to a new place (e.g., a new city or town) for work, but then every few days moved around among specific locations within that city or town, the move to that city or town was counted as a single move; however, if the respondent moved to a different city or town, it was counted as a separate move even if the respondent stayed there less than a month.

[8] Sequence analysis identifies the common patterns by minimizing within-cluster differences and maximizing between-cluster differences. Although there are many unique migration trajectories, each trajectory is clustered with others that share the most similarity, and, by doing so, we are able to identify typical trajectories.

DATA AVAILABILITY STATEMENT

The data upon which this analysis is based are currently not publicly available. The third author, Donald Treiman, has been preparing the data set for public release and is nearly finished. It will be publicly available soon. Please check his web page for availability (https://ccpr.ucla.edu/dtreiman/).

ORCID

Zhenxiang Chen https://orcid.org/0000-0001-9648-3542
Yao Lu https://orcid.org/0000-0002-6715-4640

REFERENCES

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, *16*(3), 471–494. https://doi.org/10.2307/204500

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology*, *96*(1), 144–185. https://doi.org/10.1086/229495

Adams, R. H. Jr., & Cuecuecha, A. (2013). The impact of remittances on investment and poverty in Ghana. *World Development*, *50*, 24–40. https://doi.org/10.1016/j.worlddev.2013.04.009

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods & Research*, *38*(3), 420–462. https://doi.org/10.1177/0049124109357532

Alderman, H., Behrman, J. R., Kohler, H. P., Maluccio, J. A., & Watkins, S. C. (2001). Attrition in longitudinal household survey data: Some tests for three developing-country samples. *Demographic Research*, *5*(4), 79–124. https://doi.org/10.4054/DemRes.2001.5.4

Angel, S., Disslbacher, F., Humer, S., & Schnetzer, M. (2019). What did you really earn last year? Explaining measurement error in survey income data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(4), 1411–1437. https://doi.org/10.1111/rssa.12463

Assaad, R., Krafft, C., & Yassin, S. (2018). Comparing retrospective and panel data collection methods to assess labor market dynamics. *IZA Journal of Development and Migration*, *8*(17), 1–34.

Beckett, M., Da Vanzo, J., Sastry, N., Panis, C., & Peterson, C. (2001). The quality of retrospective data: An examination of long-term recall in a developing country. *Journal of Human Resources*, *36*(3), 593–625. https://doi.org/10.2307/3069631

Bell, M., Bernard, A., Charles-Edwards, E., & Zhu, Y. (Eds.) (2020). *Internal migration in the countries of Asia: A cross-national comparison*. Springer. https://doi.org/10.1007/978-3-030-44010-7

Bell, M., Charles-Edwards, E., Ueffing, P., Stillwell, J., Kupiszewski, M., & Kupiszewska, D. (2015). Internal migration and development: Comparing migration intensities around the world. *Population and Development Review*, *41*(1), 33–58. https://doi.org/10.1111/j.1728-4457.2015.00025.x

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, *6*(4), 383–406. https://doi.org/10.1080/741942610

Bernard, A. (2017). Levels and patterns of internal migration in Europe: A cohort perspective. *Population Studies*, *71*(3), 293–311. https://doi.org/10.1080/00324728.2017.1360932

Bernard, A., Bell, M., & Charles-Edwards, E. (2014). Life-course transitions and the age profile of internal migration. *Population and Development Review*, *40*(2), 213–239. https://doi.org/10.1111/j.1728-4457.2014.00671.x

Bernard, A., Bell, M., & Zhu, Y. (2019). Migration in China: A cohort approach to understanding past and future trends. *Population, Space and Place*, *25*(6), 1–15.

Bernard, A., & Perales, F. (2021). Is migration a learned behaviour? Understanding the impact of past migration on future migration. *Population and Development Review*. https://doi.org/10.1111/padr.12387

Bernard, A., & Vidal, S. (2020). Does moving in childhood and adolescence affect residential mobility in adulthood? An analysis of long-term individual residential trajectories in 11 European countries. *Population, Space and Place*, *26*(1), 1–16.

Billari, F. C., & Piccarreta, R. (2005). Analyzing demographic life courses through sequence analysis. *Mathematical Population Studies*, *12*(2), 81–106. https://doi.org/10.1080/08898480590932287

Blane, D. B. (1996). Collecting retrospective data: Development of a reliable method and a pilot study of its use. *Social Science & Medicine*, *42*(5), 751–757. https://doi.org/10.1016/0277-9536(95)00340-1

Bourguignon, F., Fournier, M., & Gurgand, M. (2007). Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons. *Journal of Economic Surveys*, *21*(1), 174–205. https://doi.org/10.1111/j.1467-6419.2007.00503.x

Brzinsky-Fay, C., & Kohler, U. (2010). New developments in sequence analysis. *Sociological Methods & Research*, *38*(3), 359–364. https://doi.org/10.1177/0049124110363371

Cao, Z., Zheng, X., Liu, Y., Li, Y., & Chen, Y. (2018). Exploring the changing patterns of China's migration and its determinants using census data of 2000 and 2010. *Habitat International*, *82*, 72–82. https://doi.org/10.1016/j.habitatint.2018.09.006

Chan, K. W. (2001). Recent migration in China: Patterns, trends, and policies. *Asian Perspective*, *25*(4), 127–155. https://doi.org/10.1353/apr.2001.0005

Chan, K. W. (2013). China: Internal migration. In I. Ness (Ed.), *The encyclopedia of global human migration* (pp. 1–15). Hoboken: Blackwell Publishing. https://doi.org/10.1002/9781444351071.wbeghm124

Chen, A., & Coulson, N. E. (2002). Determinants of urban migration: Evidence from Chinese cities. *Urban Studies*, *39*(12), 2189–2197. https://doi.org/10.1080/0042098022000033818

Chen, J. (2013). Perceived discrimination and subjective well-being among rural-to-urban migrants in China. *Journal of Sociology and Social Welfare*, *40*, 131.

Chen, Y. (2011). Occupational attainment of migrants and local workers: Findings from a survey in Shanghai's manufacturing sector. *Urban Studies*, *48*(1), 3–21. https://doi.org/10.1177/0042098009360685

Cheung, N. W. (2013). Rural-to-urban migrant adolescents in Guangzhou, China: Psychological health, victimization, and local and trans-local ties. *Social Science & Medicine*, *93*, 121–129. https://doi.org/10.1016/j.socscimed.2013.06.021

Chunyu, M. D., Liang, Z., & Wu, Y. (2013). Interprovincial return migration in China: Individual and contextual determinants in Sichuan province in the 1990s. *Environment and Planning A*, *45*(12), 2939–2958. https://doi.org/10.1068/a45360

Constant, A. F., & Zimmermann, K. F. (2011). Circular and repeat migration: Counts of exits and years away from the host country. *Population Research and Policy Review*, *30*(4), 495–515. https://doi.org/10.1007/s11113-010-9198-6

Coulter, R., Vam Ham, M., & Findlay, A. M. (2016). Re-thinking residential mobility: Linking lives through time and space. *Progress in Human Geography*, *40*(3), 352–374. https://doi.org/10.1177/0309132515575417

Coulter, R., & Van Ham, M. (2013). Following people through time: An analysis of individual residential mobility biographies. *Housing Studies*, *28*(7), 1037–1055. https://doi.org/10.1080/02673037.2013.783903

Coulter, R., Van Ham, M., & Feijten, P. (2011). A longitudinal analysis of moving desires, expectations and actual moving behaviour. *Environment and Planning A*, *43*(11), 2742–2760. https://doi.org/10.1068/a44105

De Jong, G. F. (2000). Expectations, gender, and norms in migration decision-making. *Population Studies*, *54*(3), 307–319. https://doi.org/10.1080/713779089

Démurger, S., & Xu, H. (2011). Return migrants: The rise of new entrepreneurs in rural China. *World Development*, *39*(10), 1847–1861. https://doi.org/10.1016/j.worlddev.2011.04.027

Démurger, S., & Xu, H. (2015). Left-behind children and return migration in China. *IZA Journal of Migration*, *4*(1), 1–21.

di Belgiojoso, E. B., & Terzera, L. (2018). Family reunification—Who, when, and how? Family trajectories among migrants in Italy. *Demographic Research*, *38*, 737–772. https://doi.org/10.4054/DemRes.2018.38.28

Dubin, J. A., & McFadden, D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica: Journal of the Econometric Society*, *52*(2), 345–362. https://doi.org/10.2307/1911493

Evans, M. D., Kelley, J., Sikora, J., & Treiman, D. J. (2010). Family scholarly culture and educational success: Books and schooling in 27 nations. *Research in Social Stratification and Mobility*, *28*(2), 171–197. https://doi.org/10.1016/j.rssm.2010.01.002

Evans, M. D. R., Kelley, J., Sikora, J., & Treiman, D. J. (2015). Scholarly culture and occupational success in 31 societies. *Comparative Sociology*, *14*(2), 176–218. https://doi.org/10.1163/15691330-12341345

Falkingham, J., Sage, J., Stone, J., & Vlachantoni, A. (2016). Residential mobility across the life course: Continuity and change across three cohorts in Britain. *Advances in Life Course Research*, *30*, 111–123. https://doi.org/10.1016/j.alcr.2016.06.001

Fan, C. C. (2000). Migration and gender in China. In C. M. Lau & J. Shen (Eds.), *China review* (pp. 423–454). Hong Kong: Chinese University Press.

Fan, C. C. (2004). Gender differences in Chinese migration. In C. Hsieh & M. Lu (Eds.), *Changing China: A geographic appraisal* (pp. 243–268). Boulder: Westview.

Fan, C. C. (2005). Modeling interprovincial migration in China, 1985-2000. *Eurasian Geography and Economics*, *46*(3), 165–184. https://doi.org/10.2747/1538-7216.46.3.165

Fan, C. C., & Huang, Y. (1998). Waves of rural brides: Female marriage migration in China. *Annals of the Association of American Geographers*, *88*(2), 227–251. https://doi.org/10.1111/1467-8306.00092

Fan, C. C., & Li, L. (2002). Marriage and migration in transitional China: A field study of Gaozhou, western Guangdong. *Environment and Planning A*, *34*(4), 619–638. https://doi.org/10.1068/a34116

Fan, S., & Zhang, X. (2004). Infrastructure and regional economic development in rural China. *China Economic Review*, *15*(2), 203–214. https://doi.org/10.1016/j.chieco.2004.03.001

Fasang, A. E., & Liao, T. F. (2014). Visualizing sequences in the social sciences: Relative frequency sequence plots. *Sociological Methods & Research*, *43*(4), 643–676. https://doi.org/10.1177/0049124113506563

Favell, A. (2011). *Eurostars and Eurocities: Free movement and mobility in an integrating Europe*. Medford, MA: John Wiley & Sons.

Gabadinho, A., Ritschard, G., Mueller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, *40*(4), 1–37.

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Ganzeboom, H. B., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research*, *25*(3), 201–239. https://doi.org/10.1006/ssre.1996.0010

Gruijters, R. J., & Ermisch, J. (2019). Patrilocal, matrilocal, or neolocal? Intergenerational proximity of married couples in China. *Journal of Marriage and Family*, *81*(3), 549–566. https://doi.org/10.1111/jomf.12538

Guilmoto, C. Z. (1998). Institutions and migrations. Short-term versus long-term moves in rural West Africa. *Population Studies*, *52*(1), 85–103. https://doi.org/10.1080/0032472031000150196

Guven, C., & Islam, A. (2015). Age at migration, language proficiency, and socioeconomic outcomes: Evidence from Australia. *Demography*, *52*(2), 513–542. https://doi.org/10.1007/s13524-015-0373-6

Halpin, B. (2014). Three narratives of sequence analysis. In P. Blanchard, F. Bühlmann, & J.-A. Gauthier (Eds.), *Advances in sequence analysis: Theory, method, applications* (pp. 75–103). New York, NY: Springer. https://doi.org/10.1007/978-3-319-04969-4_5

Hennig, C., & Liao, T. F. (2010). *Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification (research report no. 308)*. London, UK: Department of Statistical Science, University College London.

Horowitz, J., & Entwisle, B. (2018). Life course and migration: A comparative case analysis of internal and international migration [paper presentation]. Population Association of America (PAA) Annual Meeting, Denver, CO.

Hu, F., Xu, Z., & Chen, Y. (2011). Circular migration, or permanent stay? Evidence from China's rural–urban migration. *China Economic Review*, *22*(1), 64–74. https://doi.org/10.1016/j.chieco.2010.09.007

Huang, Y. (2001). Gender, hukou, and the occupational attainment of female migrants in China (1985–1990). *Environment and Planning A*, *33*(2), 257–279. https://doi.org/10.1068/a33194

Impicciatore, R., & Panichella, N. (2019). Internal migration trajectories, occupational achievement and social mobility in contemporary Italy. A life course perspective. *Population, Space and Place*, *25*(6), 1–19.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). New York, NY: John Wiley & Sons.

Kim, C., & Tamborini, C. R. (2014). Response error in earnings: An analysis of the survey of income and program participation matched with administrative data. *Sociological Methods & Research*, *43*(1), 39–72. https://doi.org/10.1177/0049124112460371

Kimbro, R. T. (2009). Acculturation in context: Gender, age at migration, neighborhood ethnicity, and health behaviors. *Social Science Quarterly*, *90*(5), 1145–1166. https://doi.org/10.1111/j.1540-6237.2009.00651.x

King, R. (2002). Towards a new map of European migration. *International Journal of Population Geography*, *8*(2), 89–106. https://doi.org/10.1002/ijpg.246

Li, H., & Zahniser, S. (2002). The determinants of temporary rural-to-urban migration in China. *Urban Studies*, *39*(12), 2219–2235. https://doi.org/10.1080/0042098022000033836

Li, X., Stanton, B., Fang, X., & Lin, D. (2006). Social stigma and mental health among rural-to-urban migrants in China: A conceptual framework and future research needs. *World Health & Population*, *8*(3), 14–31. https://doi.org/10.12927/whp.2006.18282

Liang, Z. (2001). The age of migration in China. *Population and Development Review*, *27*(3), 499–524. https://doi.org/10.1111/j.1728-4457.2001.00499.x

Liang, Z. (2004). Patterns of migration and occupational attainment in contemporary China: 1985–1990. *Development and Society*, *33*(2), 251–274.

Liang, Z., & White, M. J. (1996). Internal migration in China, 1950–1988. *Demography*, *33*(3), 375–384. https://doi.org/10.2307/2061768

Liao, T. F., & Gan, R. Y. (2020). Filipino and Indonesian migrant domestic workers in Hong Kong: Their life courses in migration. *American Behavioral Scientist*, *64*(6), 740–764. https://doi.org/10.1177/0002764220910229

Lu, Y., & Qin, L. (2014). Healthy migrant and salmon bias hypotheses: A study of health and internal migration in China. *Social Science & Medicine*, *102*, 41–48. https://doi.org/10.1016/j.socscimed.2013.11.040

Lu, Y., & Treiman, D. J. (2008). The effect of sibship size on educational attainment in China: Period variations. *American Sociological Review*, *73*(5), 813–834. https://doi.org/10.1177/000312240807300506

Luo, R., Zhang, L., Huang, J., & Rozelle, S. (2007). Elections, fiscal reform and public goods provision in rural China. *Journal of Comparative Economics*, *35*(3), 583–611. https://doi.org/10.1016/j.jce.2007.03.008

Ma, Z. (2002). Social-capital mobilization and income returns to entrepreneurship: The case of return migration in rural China. *Environment and Planning A*, *34*(10), 1763–1784. https://doi.org/10.1068/a34193

MacIndoe, H., & Abbott, A. (2004). Sequence analysis and optimal matching techniques for social science data. In A. Bryman & M. Hardy (Eds.), *Handbook of data analysis* (pp. 387–406). London, UK: Sage Publications. https://doi.org/10.4135/9781848608184.n17

Meng, L., & Zhao, M. Q. (2018). Permanent and temporary rural–urban migration in China: Evidence from field surveys. *China Economic Review*, *51*, 228–239. https://doi.org/10.1016/j.chieco.2017.10.001

Meng, X. (1998). Gender occupational segregation and its impact on the gender wage differential among rural-urban migrants: A Chinese case study. *Applied Economics*, *30*(6), 741–752. https://doi.org/10.1080/000368498325444

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. https://doi.org/10.1007/BF02294245

Moore, J. C., Stinson, L. L., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics-Stockholm*, *16*(4), 331–362.

Mulder, C. H. (1993). *Migration dynamics: A life course approach*. Amsterdam: Thesis Publisher.

National Bureau of Statistics of the People's Republic of China. (2008). *Statistical division of urban and rural areas*. Beijing: Statistics Press.

National Bureau of Statistics of the People's Republic of China. (2019). *Yearbook 2019*. Beijing: Statistics Press.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Ou, D., & Kondo, A. (2013). In search of a better life: The occupational attainment of rural and urban migrants in China. *Chinese Sociological Review*, *46*(1), 25–59. https://doi.org/10.2753/CSA2162-0555460102

Parvathi, P., & Waibel, H. (2016). Organic agriculture and fair trade: A happy marriage? A case study of certified smallholder black pepper farmers in India. *World Development*, *77*, 206–220. https://doi.org/10.1016/j.worlddev.2015.08.027

Qin, L., Chen, C. P., Liu, X., Wang, C., & Jiang, Z. (2015). Health status and earnings of migrant workers from rural China. *China & World Economy*, *23*(2), 84–99. https://doi.org/10.1111/cwe.12108

Schoumaker, B. (2014). *Quality and consistency of DHS fertility estimates, 1990 to 2012*. Rockville, Maryland: ICF International.

Shen, J. (2012). Changing patterns and determinants of interprovincial migration in China 1985–2000. *Population, Space and Place*, *18*(3), 384–402.

Smith, J. P., & Thomas, D. (2003). Remembrances of things past: Test–retest reliability of retrospective migration histories. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *166*(1), 23–49. https://doi.org/10.1111/1467-985X.00257

Stovel, K., & Bolan, M. (2004). Residential trajectories: Using optimal alignment to reveal the structure of residential mobility. *Sociological Methods & Research*, *32*(4), 559–598. https://doi.org/10.1177/0049124103262683

Studer, M. (2013). *Weighted Cluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R (LIVES working papers, 24)*. Geneva, Switzerland: University of Geneva Institute for Demographic and Life Course Studies.

Tegegne, A. D., & Penker, M. (2016). Determinants of rural out-migration in Ethiopia: Who stays and who goes? *Demographic Research*, *35*(34), 1011–1044. https://doi.org/10.4054/DemRes.2016.35.34

Treiman, D. J. (2007). Appendix D: Sample design specifications. In D. J. Treiman (Ed.), *Codebook for the 2008 survey: Internal migration and health in China (IMHC)*. Los Angeles: University of California at Los Angeles, California Center for Population Research.

Treiman, D. J. (2012). The "difference between heaven and earth": Urban–rural disparities in well-being in China. *Research in Social Stratification and Mobility*, *30*(1), 33–47. https://doi.org/10.1016/j.rssm.2011.10.001

Vidal, S., & Lutz, K. (2018). Internal migration over young adult life courses: Continuities and changes across cohorts in West Germany. *Advances in Life Course Research*, *36*, 45–56. https://doi.org/10.1016/j.alcr.2018.03.003

Wang, W. W., & Fan, C. C. (2006). Success or failure: Selectivity and reasons of return migration in Sichuan and Anhui, China. *Environment and Planning A*, *38*(5), 939–958. https://doi.org/10.1068/a37428

Wingens, M., De Valk, H., Windzio, M., & Aybek, C. (2011). The sociological life course approach and research on migration and integration. In *A life-course perspective on migration and integration* (pp. 1–26). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1545-5_1

Wu, X., & Treiman, D. J. (2004). The household registration system and social stratification in China: 1955–1996. *Demography*, *41*(2), 363–384. https://doi.org/10.1353/dem.2004.0010

Wu, X., & Treiman, D. J. (2007). Inequality and equality under Chinese socialism: The *hukou* system and intergenerational occupational mobility. *American Journal of Sociology*, *113*(2), 415–445. https://doi.org/10.1086/518905

Wu, Z., & Yao, S. (2003). Intermigration and intramigration in China: A theoretical and empirical analysis. *China Economic Review*, *14*(4), 371–385. https://doi.org/10.1016/j.chieco.2003.08.001

Yan, X., Bauer, S., & Huo, X. (2014). Farm size, land reallocation, and labour migration in rural China. *Population, Space and Place*, *20*(4), 303–315. https://doi.org/10.1002/psp.1831

Yang, M., Dijst, M., & Helbich, M. (2020). Migration trajectories and their relationship to mental health among internal migrants in urban China: A sequence alignment approach. *Population, Space and Place*, *26*(5), 1–11.

Yang, X. (2000). Determinants of migration intentions in Hubei province, China: Individual versus family migration. *Environment and Planning A*, *32*(5), 769–787. https://doi.org/10.1068/a32114

Yang, X., & Guo, F. (1999). Gender differences in determinants of temporary labor migration in China: A multilevel analysis. *International Migration Review*, *33*(4), 929–953. https://doi.org/10.1177/019791839903300405

Zhang, K. H., & Song, S. (2003). Rural–urban migration and urbanization in China: Evidence from time-series and cross-section analyses. *China Economic Review*, *14*(4), 386–400. https://doi.org/10.1016/j.chieco.2003.09.018

Zhang, Z., & Treiman, D. J. (2013). Social origins, hukou conversion, and the wellbeing of urban residents in contemporary China. *Social Science Research*, *42*(1), 71–89. https://doi.org/10.1016/j.ssresearch.2012.08.004

Zhao, Y. (2003). The role of migrant networks in labor migration: The case of China. *Contemporary Economic Policy*, *21*(4), 500–511. https://doi.org/10.1093/cep/byg028

Zheng, Z., & Yang, G. (2016). Internal migration in China: Changes and trends. In C. Z. Guilmoto & G. W. Jones (Eds.), *Contemporary demographic transformations in China, India and Indonesia* (pp. 223–237). New York, NY: Springer. https://doi.org/10.1007/978-3-319-24783-0_14

Zhu, N. (2002). The impacts of income gaps on migration decisions in China. *China Economic Review*, *13*(2–3), 213–230. https://doi.org/10.1016/S1043-951X(02)00074-3

Zufferey, J., Steiner, I., & Ruedin, D. (2021). The many forms of multiple migrations: Evidence from a sequence analysis in Switzerland, 1998 to 2008. *International Migration Review*, *55*(1), 254–279. https://doi.org/10.1177/0197918320914239

How to cite this article: Chen, Z., Lu, Y., & Treiman, D. J. (2022). Determinants and consequences of rural-to-urban migration patterns in China: Evidence from sequence analysis. *Population, Space and Place*, 28, e2493. https://doi.org/10.1002/psp.2493

APPENDIX A: COMPARING IMHC AND CHARLS

Among all other publicly available surveys in China, only The China Health and Retirement Longitudinal Study (CHARLS) includes high-quality migration history data. CHARLS began in 2011 with a nationally representative sample of about 17,500 individuals age 45 and older and has collected data in four waves so far, the most recent in 2018. A detailed migration history was collected in 2014, when respondents were age 48 and older. While a large sample of individuals who have essentially completed their labor migration experience would seem to be an advantage, there are several important limitations to the CHARLS data. One is that for all respondents, the bulk of their labor migration occurred more than 20 years before they were queried, raising important questions about recall bias of the sort discussed above. But there also are several features of the way migration histories were collected that make the CHARLS data less desirable for our purposes.

First, compared with CHARLS, IMHC has a much more detailed measure of migration. In IMHC, each place where the respondent stayed for more than a month was treated as a new episode, which means that the survey captures both movement from one migration destination to another and episodes of return migration. For each episode, information was collected on both the characteristics of that place and characteristics of the respondent at the time (details below). It also includes several restrictions[7] to prevent miscounting. In comparison, CHARLS defines changes of residence using a six month threshold, which misses some migration episodes.

In addition, IMHC has a richer locational information that captures all 10 potential levels of each place the respondent migrated, including "small village (<1,000)," "ordinary village (1,000–2,500)," "large village (>2,500)," "township seat of a xiang," "township seat of a zhen," "county seat," "county-level city," "prefecture-level city," "province capital," and "province-level city." The most detailed locational information is at county level in CHARLS. Third, in CHARLS, rural and urban differentiation is obtained by asking the respondents directly instead of being based on locational information. But it is hard to use information like "mainly living in the rural area," "mainly living in the urban area," or "time living in rural and urban roughly the same" to define rural versus urban.

TABLE A1 Migration frequency by type

	Percentage/mean
For rural-to-urban migrants and non-migrants	
Number of migrations	
0	40.65
1	19.33
2	19.51
3 or more	20.51
N	1,112
For rural-to-urban migrants	
Average number of migrations	2.38
Number of rural migrations/number of rural destinations	
0	66.97/75.76
1	19.09/18.03
2	9.55/4.85
3	2.12/1.21
4 or more	2.27/0.15
Average number of rural migrations/number of rural destinations	0.57/0.32
Number of urban migrations/number of urban destinations	
0	0.00/0.00
1	52.88/59.09
2	27.58/26.36
3	11.82/10.00
4 or more	7.73/4.55
Average number of urban migrations/number of urban destinations	1.81/1.63
Number of return migrations	
0	81.52
1	15.30
2	2.27
3 or more	0.91
Average number of return migrations	0.23
N	660

APPENDIX B: SEQUENCE ANALYSIS PROCEDURES

First, we assign a migration state at each age and constructed a migration trajectory (sequence) for each respondent. Second, we calculate dissimilarities between sequences using the optimal matching algorithm (MacIndoe & Abbott 2004; Needleman & Wunsch 1970). The total

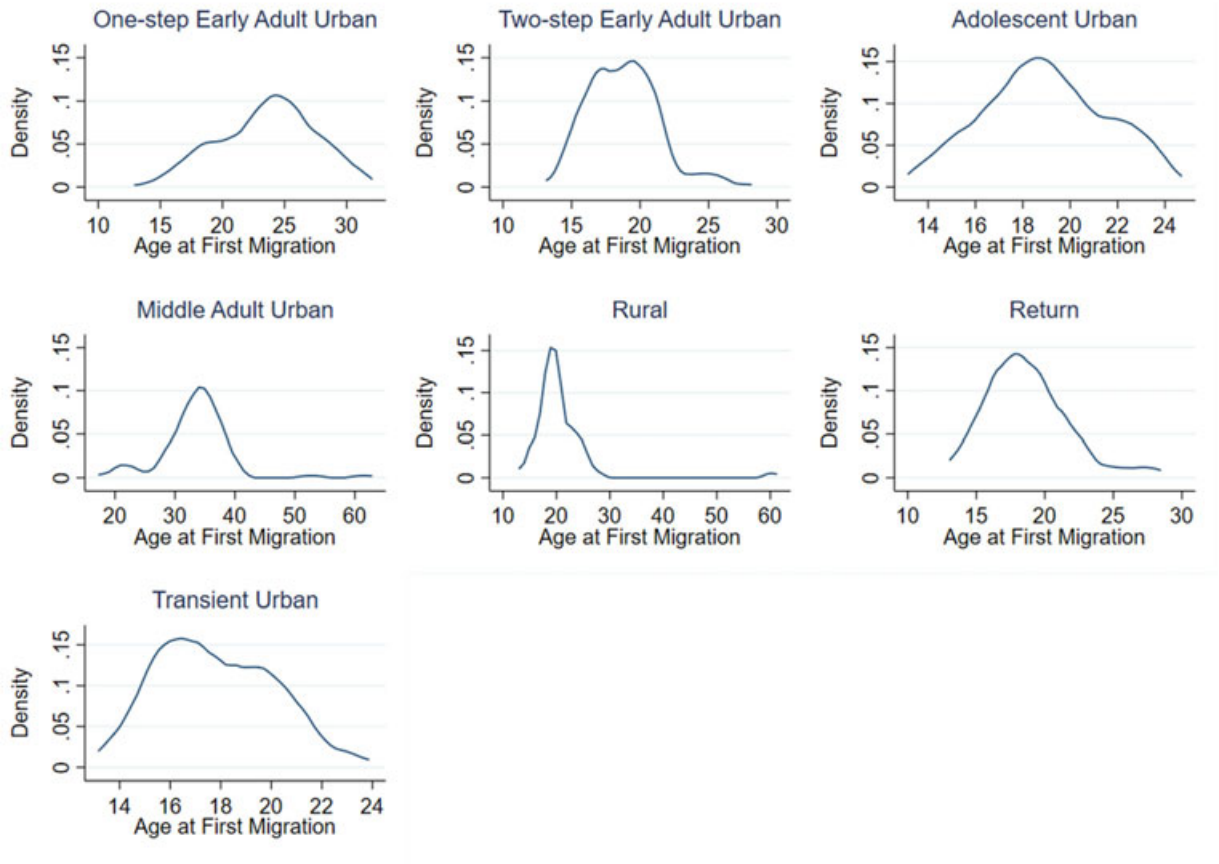

FIGURE A1 Kernel density plots for age at first migration by migration patterns

"cost" of turning one sequence into another is determined by the similarity between two sequences (Brzinsky-Fay & Kohler 2010). We use transition rates (probability) between two states as the substitution costs (Halpin 2014) and used the default indel cost (insertion and deletion cost of 1). This yields a dissimilarity matrix for every pair of sequences in the data. We explore alternative ways to calculate total "costs," including (a) using the time-varying transition matrix as substitution costs and (b) using different indel costs. The results are consistent.

In the last step, we subject the dissimilarity matrix to cluster analysis and identified common migration patterns from a large number of sequences. In particular, our cluster analysis is based on Ward's hierarchical fusion algorithm (Hennig & Liao 2010; Kaufman & Rousseeuw 2009; Milligan & Cooper 1985) combined with the Partitioning Around Medoids (PAM) algorithm. The combined approach maintains the advantages of each algorithm and allows for optimizing a global criterion. We use results from the hierarchical clustering procedure to initialize the PAM algorithm medoids. Ward's hierarchical fusion algorithm and PAM were also conducted separately and their corresponding Average Silhouette Width Weighted ("ASWw"), Hubert's Gamma ("HG"), Point Biserial Correlation ("PBC"), and Hubert's C ("HC") indexes were used to identify the optimal number of distinctive patterns where each sequence fits into one of these patterns as closely as possible.[8] Based on these procedures, we reach a seven-cluster solution.

APPENDIX C: DMF ESTIMATION PROCEDURE AND ATT ESTIMATES

Consider the following model:

$$y_1 = x\beta_1 + \mu_1,$$
$$y_j^* = z\gamma_j + \theta_j, j = 1,...,8,$$

where y_1 is the observed outcome if migrant choose pattern 1, vector x contains all determinants of the variable of interest, and μ_1 is the disturbance term with $E(\mu_1|\, x,z) = 0$ and $Var(\mu_1|\, x,z) = \sigma$, where j is a categorical variable that describes the choice of a migrant among eight alternatives based on latent outcome y_j^*, vector z contains explanatory variables for all alternatives, and θ_j is the disturbance term.

Here, y_1 is observed if and only if pattern 1 is chosen. Pattern 1 is chosen when

$$y_1^* > \max_{j \neq 1}\left(y_j^*\right),$$
$$\max_{j \neq 1}\left(y_j^* - y_1^*\right) < 0,$$
$$\max_{j \neq 1}\left(z\gamma_j + \theta_j - z\gamma_1 - \theta_1\right) < 0.$$

Assume that the (θ_j)s are independent and identically Gumbel distributed. Then,

$$P(Pattern\,1\,is\,chosen|z) = \frac{\exp(z\gamma_1)}{\sum_j \exp(z\gamma_j)}.$$

The problem, however, is that μ_1 maybe correlated with some (θ_j) s, in which case the estimation of β_1 will not be consistent with OLS regression.

Following the Dubin and McFadden (1984) linearity assumption,

$$E(\mu_1|\theta_1 \ldots \theta_8) = \sigma \frac{\sqrt{6}}{\pi} \sum_{j=1\ldots8} \rho_j (\theta_j - E(\theta_j)),$$

where ρ_j is the correlation between μ_1 and θ_j.

Now, define τ as follows:

$$\tau = [z\gamma_1, \ldots, z\gamma_8].$$

Let P_k be the probability that any alternative k is preferred:

$$P_k = \frac{exp(z\gamma_k)}{\sum_j exp(z\gamma_j)}.$$

We also have the following based on the multinomial logit model:

$$E\left(\theta_1 - E(\theta_1)|y_1^* > \max_{s \neq 1}(y_s^*), \tau\right) = -\ln(P_1),$$
$$E\left(\theta_j - E(\theta_j)|y_1^* > \max_{s \neq 1}(y_s^*), \tau\right) = \frac{P_j \ln(P_j)}{1 - P_j}, \forall j > 1.$$

Then, β_1 can be estimated by the following OLS regression:

$$y_1 = x\beta_1 + \sigma \frac{\sqrt{6}}{\pi} \left[\sum_{j=2\ldots8} \rho_j \left(\frac{P_j \ln(P_j)}{1 - P_j} \right) - \rho_1 \ln(P_1) \right] + \omega_1,$$

where ω_1 is a residual that is mean-independent of the regressors.

Based on the coefficients obtained, we can estimate ATT in the following way:

$$ATT(\mathrm{j}) = \mathrm{E}(ISEI_{ij}|Pattern_i = \mathrm{j}) - \mathrm{E}(ISEI_{istay}|Pattern_i = \mathrm{j}),$$

where $\mathrm{E}(ISEI_{ij}|\ Pattern_i = \mathrm{j}) = f(x_i\beta_j,\ \varphi_j\rho_j)$ while $\mathrm{E}(ISEI_{istay}|\ Pattern_i = \mathrm{j}) = f(x_i\beta_{stay},\ \varphi_j\rho_{stay})$. Here, φ is the inverse mills ratio.

APPENDIX D: SENSITIVITY ANALYSES

D.1 | Right censoring

For those who are younger than age 40 at the time of the survey, we do not observe their complete migration histories in the 14- to 40-observation window. That is, these observations are right-censored. As a sensitivity check, we conduct the sequence analysis for only those respondents who were at least 40 years old at the time of the survey. Figures A2 and A3 display the sequence analysis results based on this restriction. We observe the exact same seven patterns as the main analysis. The fact that the restriction to age 40 yields essentially the same results suggests that right censoring is not a serious issue.

D.2 | Reducing the number of clusters

To reduce the number of clusters in order to minimize the number of clusters that include only a small number of cases, one option is to choose an alternative solution based on Ward's hierarchical fusion algorithm and the Partitioning Around Medoids (PAM) algorithm.A five-cluster solution is also acceptable. If we did this, the original Cluster 5 ("Return") and Cluster 7 ("Transient Urban") would be subsumed by other clusters. However, Cluster 6 ("Rural") would remain the same and thus we would still have a small sample size for this group of migrants. Instead, a more feasible and possibly better solution would be to merge some clusters based on their characteristics. Cluster 5 ("Return") and Cluster 6 ("Rural") could be grouped together to form a new Cluster 5N ("Rural") since their home at the time of the survey is rural. Similarly, Cluster 7 ("Transient Urban") could be grouped with Cluster 2 ("Two-Step Early Adult Urban") to form a new Cluster 2N ("Multiple-Step Early Adult Urban").

We conduct the same set of analyses based on these new clusters. Multinomial logit regression results are shown in Table A2. Comparing Table A2 with Table 3, we observe the same determinants for Clusters 1, 3, and 4 (i.e., the three original clusters). Although we have two new clusters—Clusters 2N and 5N—we reach similar conclusions as in our initial analysis. In particular, if we compare Cluster 2 with Cluster 2N, the only difference is that communities with more types of schools increase the likelihood of being in Cluster 2N but not Cluster 2. Therefore, the conclusion would be that communities with more types of schools tend to produce migrants who moved to multiple, instead of three or more, places. Similarly, comparing Cluster 5N with Clusters 5 and 6, we would observe different determinants but similar conclusions. The conclusion that rural-to-urban migration increase was mainly driven by those who migrated as young adults would include those who already had returned but also those who had moved to other rural areas. Also, initially we found that women were much more likely than men to be included in clusters (1 and 3) that were driven by marriage migration (see Table 1). They also were more likely than men to be found in Cluster 5N, but only at the 10% significance level.

The analysis results on the consequences of migration patterns are shown in Table A3. Comparing Table A3 with Table 6, we reach similar conclusions for both ISEI and ISEI slope. The main difference is that adult urban migration, instead of non-transient adult urban migration, is associated with better occupational outcomes, measured by both occupational status and occupational trajectories.

To conclude, the overall difference between the original seven-cluster solution and the five-cluster solution is minimum. Except for

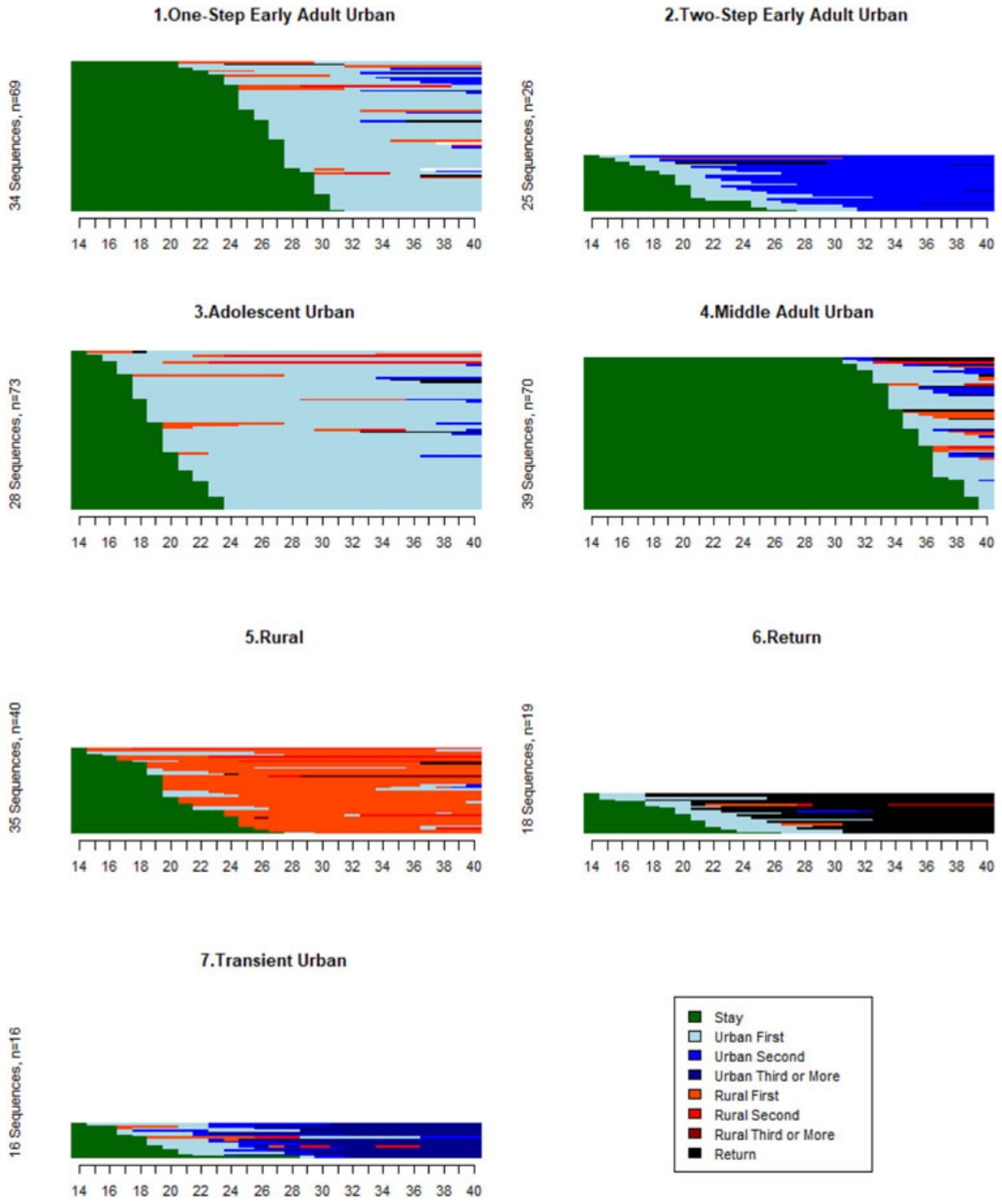

FIGURE A2 Sequence index plots of migration trajectory clusters (for migrants older than 40), IMHC 2008 (weighted frequencies are shown)

the differences mentioned above, all the rest of the main findings hold between the two solutions. However, while we reach mostly similar conclusions, we tend to lose information using the five-cluster solution (especially the difference between rural and return pattern and the uniqueness of the transient migration). Therefore, we deem the seven-cluster solution superior and use it for the main analyses.

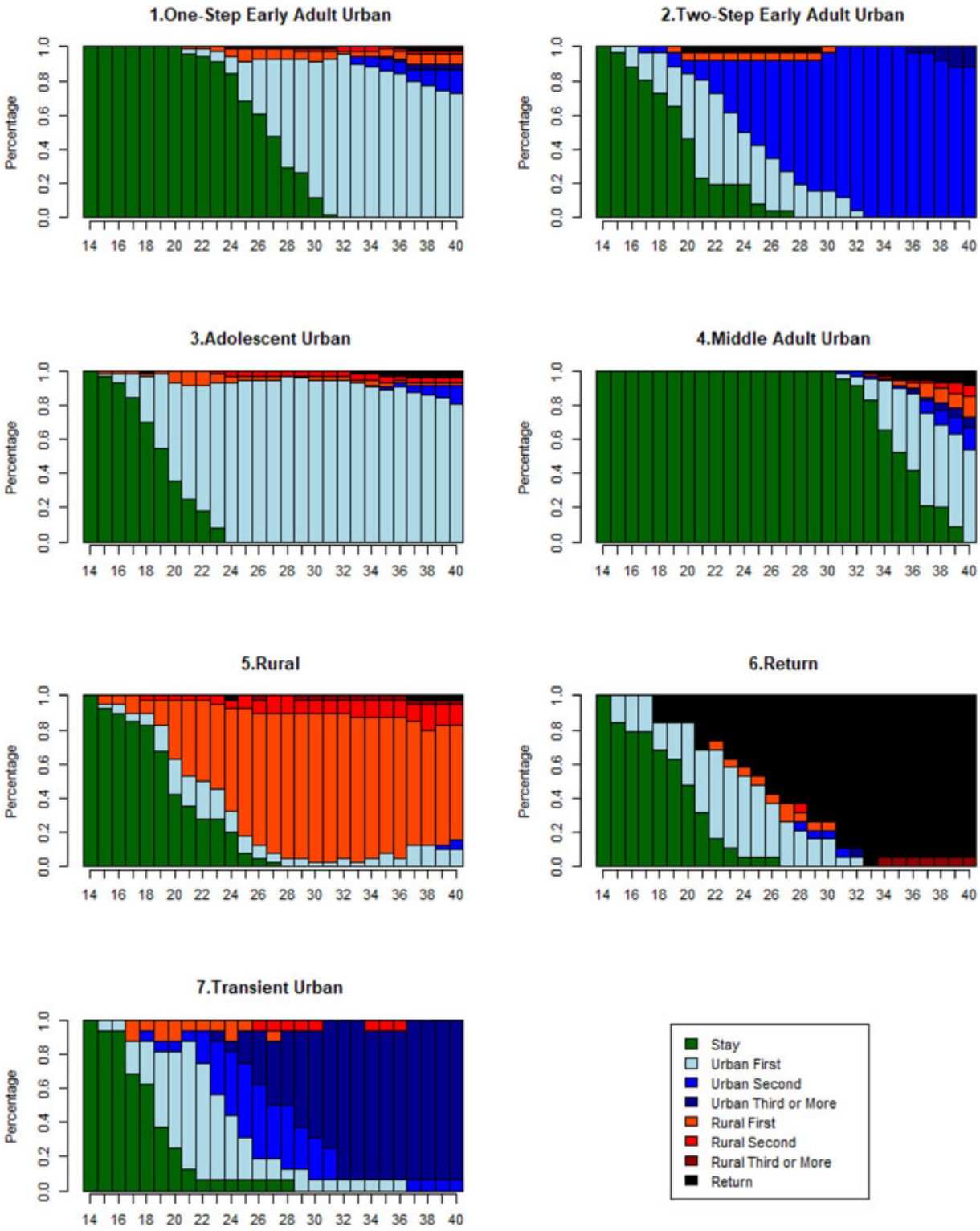

FIGURE A3 State distribution plots by migration trajectory clusters (for migrants older than 40), IMHC 2008

TABLE A2 Multinomial logistic regression of determinants of migration patterns (merged)

	Patterns				
	One-step early adult urban	Multiple-steps early adult urban	Early adolescent urban	Middle adult urban	Rural
VARIABLES	Same	Two-steps early Adult Urban & Transient Urban	Same	Same	Rural & Return
Birth year	0.070*** (0.011)	0.066*** (0.012)	0.003 (0.012)	0.008 (0.012)	0.047*** (0.012)
Female	0.848*** (0.201)	0.122 (0.210)	0.800*** (0.226)	0.136 (0.239)	0.443^{+} (0.227)
Education (ref = no school)					
Less than primary	−0.333 (0.469)	0.483 (0.835)	−0.458 (0.548)	0.790 (0.553)	0.637 (0.522)
Primary	−0.443 (0.439)	0.483 (0.791)	0.307 (0.467)	0.389 (0.551)	0.704 (0.505)
Middle or more	0.106 (0.420)	1.301^{+} (0.770)	0.494 (0.460)	0.394 (0.545)	0.307 (0.511)
Book at 14 (ref = 0)					
1–9	−0.316 (0.283)	0.093 (0.362)	−0.036 (0.326)	−0.136 (0.356)	−0.350 (0.311)
10–19	−0.299 (0.315)	0.525 (0.370)	0.429 (0.344)	0.895* (0.359)	0.137 (0.338)
20+	−0.226 (0.316)	0.606 (0.370)	−0.071 (0.382)	0.960* (0.376)	0.075 (0.357)
Father's education (ref = no school)					
Less than primary	0.490^{+} (0.274)	0.087 (0.308)	0.183 (0.294)	0.038 (0.298)	−0.212 (0.292)
Primary	0.714* (0.304)	0.850** (0.311)	0.404 (0.332)	0.332 (0.334)	−0.177 (0.346)
Middle or more	0.490 (0.343)	0.252 (0.360)	0.079 (0.394)	−0.830^{+} (0.499)	−0.039 (0.385)
Father read (ref = does not read)	−0.134 (0.237)	−0.160 (0.245)	0.303 (0.269)	−0.495 (0.315)	−0.359 (0.302)
Protein intake at age 14	0.536* (0.213)	0.774*** (0.223)	0.338 (0.255)	0.709** (0.259)	0.309 (0.254)
Home distance to the county seat (ref = half a day)					
More than a half day	0.389 (0.286)	0.746** (0.289)	0.486 (0.316)	0.399 (0.321)	−0.173 (0.344)
In county seat	−0.174 (0.252)	−0.362 (0.271)	0.115 (0.271)	−0.595^{+} (0.344)	−1.379*** (0.417)
Number of school types (ref = primary only)					
Primary and middle	−0.161 (0.307)	0.615^{+} (0.350)	−0.230 (0.320)	−0.060 (0.329)	−0.237 (0.329)
Primary, middle, and high	0.265 (0.255)	0.835** (0.308)	−0.187 (0.270)	−0.059 (0.288)	0.084 (0.279)
Constant	−138.532*** (22.184)	−132.811*** (23.705)	−8.334 (23.176)	−17.303 (23.963)	−94.088*** (23.944)
Observations	1,112	1,112	1,112	1,112	1,112

Note: Standard errors in parentheses. Coefficients are logits (log-odds); stayers—non-migrants—is the reference category.
***p < 0.001, **p < 0.01, *p < 0.05. ^{+}p < 0.1.

TABLE A3 Average treatment effects on the treated (ATT) for the five-migration category specification

ISEI	Migration patterns		Estimated	Counterfactual	ATT	p value
Current	One-step early adult urban	DMF	40.61	35.32	5.29	0.000
	Multiple-step early adult urban	DMF	41.41	37.50	3.91	0.000
	Adolescent urban	DMF	36.35	36.09	0.26	0.702
	Middle adult urban	DMF	35.56	32.99	2.57	0.003
	Rural	DMF	32.04	31.71	.326	0.563
Trajectory	One-step early adult urban	DMF	0.248	0.106	0.142	0.000
	Multiple-step early adult urban	DMF	0.262	0.154	0.108	0.000
	Adolescent urban	DMF	0.075	0.087	−0.011	0.654
	Middle adult urban	DMF	0.193	0.089	0.103	0.004
	Rural	DMF	0.128	0.097	0.031	0.517