Migrants and Firms: Evidence from China

Clement Imbert Marlon Seror Yifan Zhang
Yanos Zylberberg*

Online Appendix

A Measures of production and labor cost in cities **2**
A.1 A census of large manufacturing firms . . . 2
A.2 A text-based classification of products . . . 5
A.3 A survey of urban workers . . . 8

B Migration flows: construction and description **10**
B.1 Elements of context . . . 10
B.2 Data sources and construction of migration flows . . . 11
B.3 Migration patterns and the selection of migrants. . . . 15

C A shift-share instrument for migration flows **19**
C.1 A stylized model . . . 19
C.2 Shocks to rural livelihoods . . . 22
C.3 Sensitivity analysis and robustness checks . . . 24

D Migration and factor productivity **30**
D.1 Quantitative framework . . . 30
D.2 Effect of migration on factor productivity . . . 32
D.3 Heterogeneous firm technology and labor efficiency . . . 34

E Additional results on heterogeneity and firm entry/exit **38**
E.1 Treatment heterogeneity across urban establishments . . . 38
E.2 Aggregation and entry/exit . . . 39
E.3 Worker heterogeneity and compositional effects at destination . . . 42
E.4 Complements on production restructuring . . . 44

F Sensitivity analysis **48**
F.1 The shift-share design . . . 48
F.2 The empirical specification . . . 48

References **56**

*Imbert: University of Warwick, BREAD, CEPR and JPAL, c.imbert@warwick.ac.uk; Seror: Université du Québec à Montréal, University of Bristol, DIAL, Institut Convergences Migrations, seror.marlon@uqam.ca; Zhang: Chinese University of Hong Kong, yifan.zhang@cuhk.edu.hk; Zylberberg: University of Bristol, CESifo, the Alan Turing Institute, yanos.zylberberg@bristol.ac.uk.

A Measures of production and labor cost in cities

In this Appendix, we describe the census of large manufacturing establishments, we detail our product classification, and we describe a survey of urban residents.

A.1 A census of large manufacturing firms

Description Our measures of urban production are derived from the census of "above-scale" manufacturing establishments conducted by the National Bureau of Statistics (NBS). The NBS implements a census of all state-owned manufacturing enterprises and all non-state manufacturing firms with sales exceeding RMB 5 million, or about $600,000 over that period.

The data cover the manufacturing sector over the period 1992–2009. The set of variables changes across years and we restrict ourselves to the period 2000–2006 in the baseline specification, in order to ensure consistent outcome measures. We focus on the balanced panel of firms in most of our analysis. In contrast with firm-level data in developed countries, matching firms over time in the NBS is difficult because of frequent changes in identifiers. In order to match "identifier-switchers," we use the fuzzy algorithm developed by Brandt et al. (2014), which uses slowly-changing firm characteristics such as name, address, and phone number. While total sample size ranges between 150,000 and 300,000 per year, we end up with about 32,000 firms when we limit the sample to the balanced panel.[1]

The data contain a wealth of information on manufacturing plants. Besides the location, industry, ownership type, exporting activity, and number of employees, they offer a wide range of accounting variables (e.g., output, input, value added, wage bill, fixed assets, financial assets, etc.). We use these variables to construct the firm-level measures of factor choices, costs, and productivity. Finally, each firm reports its three main products as a textual description that we exploit with a language processing algorithm in order to generate a consistent HS-6 product code.

Descriptive statistics and sample selection Table A1 displays descriptive statistics for the sample of all firm × year observations over the period 2000–2006, the balanced panel, and the sub-samples of new entrants and exiters. Firms of the

[1] Although we use the term "firm" in the paper, the NBS data cover "legal units" (*faren danwei*): different subsidiaries of the same enterprise may be surveyed, provided they meet a number of criteria, including having their own names, being able to sign contracts, possessing and using assets independently, assuming their liabilities, and being financially independent. While this definition of units of observation may be unfamiliar to readers accustomed to U.S. or European data, "legal units" almost perfectly overlap with plants in practice, which is also true of establishments in the U.S. In 2006, almost 97% of the units in our data corresponded to single-plant firms.

balanced panel are larger and more capitalized than the average firm (see Panel A). They are also more likely to be publicly owned.[2] The difference between the balanced panel and whole sample comes from inflows (new entrants) and outflows (exiters). The third and fourth columns of Table A1 better characterize these two categories of firms. Firms on the brink of exit are small, under-capitalized, unproductive, and less likely to be located in an industrial cluster. New entrants are equally small and under-capitalized, but comparatively productive. Finally, the period of interest is a period of public sector downsizing. While private firms still accounted for a relatively small share of the economic activity in the 1990s, they represented over 80% of total value added by the end of the 2000s. We see indeed that new entrants are disproportionately privately owned.

Issues This "above-scale" census raises a number of challenges. First, the census covers the universe of state-owned enterprises but only a selected sample of private establishments, even if together they account for about 90% of the manufacturing output. The RMB 5 million threshold that defines whether a non-publicly owned firm belongs to the NBS census may not be perfectly implemented. Surveyors may not correctly predict the level of sales before implementing the census, and some firms only entered the database several years after having reached the sales cut-off.[3] Figure 1 however shows that this is unlikely to be a major issue, as the threshold is sharp.

Second, firms may under-report the number of workers. Indeed, firm size serves as a basis for taxation by the local labor department, and migrants, who represent a large share of the workforce, may be easier to under-report. Workers hired through a "labor dispatching" (*laodong paiqian*) company are not included in the employment variable. The wage bill may also be slightly under-estimated as some components of worker compensation are not recorded in all years, e.g., pension contributions and housing subsidies, which are reported only since 2003 and 2004, respectively, but accounted for only 3.5% of total worker compensation in 2007.

Third, a few variables are not documented in the same way as in standard firm-level data. Fixed assets are reported by summing nominal values at the time of purchase. We use the procedure developed in Brandt et al. (2014) to account for depreciation: (i) We calculate the nominal rate of growth in the capital stock (using a

[2]Ownership type is defined based on official registration (*qiye dengji zhuce leixing*). Table A1 uses three categories: (i) state-owned, hybrid or collective, (ii) domestic private, and (iii) foreign private firms, including those from Hong Kong, Macau, and Taiwan.

[3]Conversely, about 5% of private and collectively owned firms, which are subject to the threshold, continue to participate in the census even if their annual sales fall short of the threshold.

Table A1. Firm characteristics (2000–2006).

	All firms	Balanced 2000–2006	Exiters	Entrants
Panel A: Outcome variables				
Labor cost	2.53	2.450	2.33	2.56
	(0.67)	(0.68)	(0.77)	(0.65)
Employment	4.64	5.09	4.19	4.47
	(1.06)	(1.06)	(1.03)	(1.03)
K/L ratio	3.68	3.90	3.50	3.54
	(1.28)	(1.18)	(1.34)	(1.32)
Value added	8.57	8.95	7.91	8.42
	(1.16)	(1.14)	(1.09)	(1.16)
Panel B: Characteristics				
Public	0.13	0.18	0.24	0.09
	(0.34)	(0.39)	(0.43)	(0.29)
Export	0.21	0.33	0.14	0.20
	(0.41)	(0.47)	(0.34)	(0.40)
Large	0.43	0.45	0.33	0.34
	(0.49)	(0.50)	(0.47)	(0.47)
High-skill	0.45	0.45	0.45	0.44
	(0.50)	(0.50)	(0.50)	(0.50)
Unionized	0.08	0.10	0.11	0.08
	(0.28)	(0.30)	(0.31)	(0.28)
Ind. park	0.13	0.11	0.10	0.16
	(0.34)	(0.31)	(0.30)	(0.37)
Observations	1,350,167	227,031	136,587	519,261

Notes: NBS firm-level data (2000–2006). Standard deviations are reported in parentheses. All variables in Panel A are in logarithms. All variables in Panel B are dummy-coded and defined for the first year in the sample. *Public* is equal to 1 if the firm is state- or collective-owned in 2000. A similar definition applies to *Export*, *Unionized*, and *Ind. park*, which are equal to 1 if the firm exported, had a trade union, and operated in an industrial park in the first year, respectively. *Large* is defined as equal to 1 if the firm belonged to the top 50% of the distribution in terms of employment. *High-skill* is equal to 1 if the firm belongs to an industry with an above-median share of tertiary-educated employees; this variable is available only in 2004 (N = 248,734, 227,031, 2,521, and 8,692 in columns 1, 2, 3, and 4, respectively).

2-digit industry by province average in 1993–1998) to compute nominal capital stock in the start-up year. (ii) Real capital in the start-up year is obtained using a chain-linked investment deflator (based on separate price indices for equipment-machinery and buildings-structures, and weighted by fixed investment shares provided by the NBS). (iii) We move forward to the first year in the database, assuming a rate of depreciation of 9% per year and using annual deflators. (iv) Once a firm enters the database, we use the nominal figures provided in the data to compute the change in nominal capital stock in a given year and deflate it. If past investments and

depreciation are not available in the data, we use information on the age of the firm and estimates of the average growth rate of nominal capital stock at the 2-digit industry level between 1993 and the year of entry in the database.

Fourth, the census of manufacturing establishments contains a Chinese 4-digit industry classification (CIC) at the establishment level,[4] but only text *descriptions* for (up to) three main products.

A.2 A text-based classification of products

This section describes how we construct a text-based classification of products in the census of manufacturing establishments.

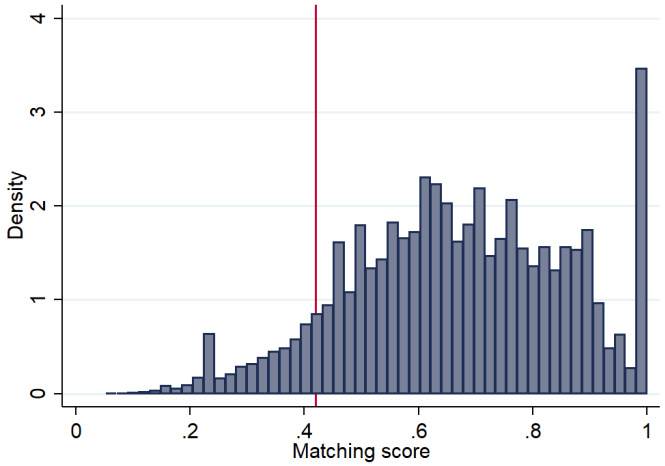
Natural language processing The text descriptions of products are not standardized, but they are often precise and specific. In principle, how they are reported is up to the firm representative who reports to the census enumerators. For these reasons, similar products can be described differently across establishments, and differently within the same establishment over time. For instance, one establishment reports "Refractory" in 2001–2003, "Production of advanced ceramic products" in 2004–2005, and "Advanced ceramic products" in 2006. Another establishment reports "Multicolor printing" in 2001–2003, and "Holiday card printing" in 2004–2006.

We develop an algorithm to match each description with a unique HS 6-digit code, exploiting the official description for these product categories in Chinese. Intuitively, the algorithm creates a similarity score between a product description in the census and these official category descriptions; the HS 6-digit category with the highest score is then associated with the product. The analysis proceeds in three steps. The first step transforms the category descriptions and the textual information provided by manufacturing firms into sequences of distinct, relevant words. A tokenizer in Chinese ("jieba") groups characters into words; we then use a list of Chinese stop words in order to filter out common words or particles. The second step consists in projecting these sequences onto a vector space—a process known as word embedding. The purpose of the projection is to capture the different dimensions of similarity between words occurring in a similar context. In the previous examples, "printing"/"holiday card" should be allocated a high semantic similarity, while "printing"/"credit card" should be represented by very different embeddings. This process is a complex problem, and we use the powerful neural net developed by

[4]There exist imperfect bridges between the CIC classification and international (or US) industry classifications. For instance, one such bridge is constructed in an influential study looking at the labor market effect of Chinese competition in the United States (Autor et al., 2013).

Google ("word2vec", see Mikolov et al., 2013) in order to represent every sequence of words in a carefully chosen vector space. The neural net produces a vector space from a training corpus, which should provide a large collection of common contexts and word associations in order to best represent semantic similarity. Ideally, we would like to train the "word2vec" model on a text describing the development of manufacturing products from production to sale. Unfortunately, the corpus needs to be very large, and we rely on the word embeddings provided by Li et al. (2018) and trained on the Wikipedia corpus. The third step is a direct application of the vector space representation. One can use the projections in the vector space in order to compute a similarity score. More specifically, we construct a normalized similarity score between 0 and 1 (for two similar sequences) from the average of the distance metric between all combinations of contiguous word sequences taken from (i) the product description and (ii) the description of a HS 6-digit code. While we collect the best 10 matches, we only use the HS 6-digit category with the highest score in our analysis. This highest score is on average quite high: about 90% of matches have a best similarity score than the similarity score that would be computed for "Multicolor printing" and "Holiday card printing" (see Figure A1).

Figure A1. Distribution of similarity scores for the main product across years (2000–2006, balanced sample of establishments).



Notes: This figure represents the distribution of similarity scores computed between the (main) product description and the best suited product category for all establishments of the balanced sample between 2000–2006. The red line represents the similarity score between "Multicolor printing" and "Holiday card printing".

Complements, match quality, and validation The output of the previous procedure is a HS 6-digit product code for each product reported by a firm. This HS 6-digit product code can be used to better characterize firm production and its

evolution during our period of interest.

We first use the HS 6-digit product code as a label capturing the specificity of a given production process. The underlying assumption is that the production process is quite homogeneous within a HS 6-digit product code, such that the relative factor use of firms producing a certain product is indicative of factor shares in production. Letting $r_{i,p}$ denote the relative factor share, e.g., capital to labor or the share of high-skilled workers, in firm i producing product p at baseline, we construct r_p as the average factor share required by the production of p. One could then characterize the direction of a change from product p to product q by comparing r_p to r_q. One issue with the previous characterization is that it only accounts for the factor share within the firm, and not possibly along the whole production line leading to the final good. We use input-output accounts in the United States (in 2002, Stewart et al., 2007) to construct a measure of factor use along the production chain, $s_p = \sum_i \alpha_{p,q} r_{i,q} \mathbb{1}_{p(i)=q} / \sum_i \alpha_{p,q} \mathbb{1}_{p(i)=q}$, where $\alpha_{p,q}$ is the contribution of product q in the production of p.

Second, we use the HS 6-digit product code to measure the production "width" of a firm: the number of different products produced by the establishment; and their similarity. We count the number of different product descriptions, but also the number of different HS product codes for different levels of disaggregation (e.g., 6-digit or 3-digit). We construct three different measures of similarity, $\gamma_{p,q}$, between products p and q: a measure based on language proximity; a measure based on production proximity; and a measure based on technological proximity. We construct language proximity as the average of similarity scores between the descriptions of unique pairs of products (only one pair if there are two products reported by the establishment, three pairs if there are three products). We construct production proximity in a similar manner using the previous input/output measure, $\alpha_{p,q}$. We construct technological proximity using a technology closeness measure, $\tau_{p,q}$, derived from patent (cross-)citations between different industries, $\{T_{i,j}\}_{i,j}$, in the United States (Bloom et al., 2013).[5]

Third, we rely on the previous technology closeness measure to construct: the intensity of technological spillovers for a given product, $\{\tau_{p,q}\}_{p,q}$; the number of links to other industries, $\sum_{q \neq p} \mathbb{1}_{\tau_{p,q}>0}$; and a Herfindahl index based on technology closeness, $\sum_{q \neq p} \tau_{p,q}^2 / (\sum_{q \neq p} \tau_{p,q})^2$. These variables capture the intensity and the "width" of technological spillovers.

[5]We need to mediate these measures through a bridge $\mathbf{B}$ between HS-6 product codes and Standard Industrial Classification (SIC) codes: $\tau = \mathbf{B}'\mathbf{T}\mathbf{B}$.

A.3 A survey of urban workers

In order to study the impact of immigration on local labor markets and isolate equilibrium effects on wages from compositional effects, we use the Urban Household Survey (UHS) collected by the National Bureau of Statistics. The UHS is a survey of urban China, with a consistent questionnaire since 1986 but considered representative from 2002 onward, and our description will correspond to this latter period. The survey is based on a three-stage stratified random sampling. Its design is similar to that of the Current Population Survey in the United States (Ge and Yang, 2014; Feng et al., 2017) and includes 18 provinces and 207 prefectures. The data are cross-sectional, with a sample size that ranges from 70,000 to 90,000 individuals in 2002–2006.[6]

The UHS is a rich dataset with detailed information on individual employment, income, and household characteristics. We construct a measure of real wages as the monthly wages divided by a prefecture- and year-specific consumer price index, which we compute using the detailed household-level consumption data. We also construct three employment outcomes: wage employment, unemployment, and self-employment (which also includes firm owners). Table A2 provides some descriptive statistics of key variables over the period 2002–2006 and shows that the sample is similar to the locally registered urban *hukou* holders in the Mini-Census data (see Table B3) in terms of demographics and sector of activity, although they tend to be more educated, have a higher probability of being employed, and earn a higher monthly income.

[6]While all households living in urban areas are eligible, sampling still ignores urban dwellers living in townships and in suburban districts (Park, 2008). Rural-urban migrants, who are more likely to live in peripheral areas of cities, are therefore underrepresented.

Table A2. Descriptive statistics on urban residents from the UHS data (2002–2006).

	Mean	Standard deviation
Age	43.1	11.0
Female	0.50	0.50
Married	0.88	0.33
Education:		
Primary education	0.05	0.21
Lower secondary	0.28	0.45
Higher secondary	0.26	0.44
Tertiary education	0.41	0.49
Not employed	0.23	0.42
Self-employed/Firm owner	0.05	0.22
Employee	0.72	0.45
...of which:		
Public sector	0.68	0.47
Private sector	0.32	0.47
Total monthly income (RMB)	1,294	1,123
Hours worked per week	44.5	9.20
Industry:		
Agriculture	0.01	0.10
Mining	0.02	0.14
Manufacturing	0.24	0.42
Utilities	0.03	0.18
Construction	0.03	0.17
Wholesale and retail trade	0.12	0.33
Other tertiary	0.42	0.49
Observations	338,221	

Notes: All variables except *Age*, *Income*, and *Hours worked per week* are dummy-coded. The table displays averages over the period 2002–2006. The sample is restricted to locally registered urban *hukou* holders aged 15–64.

B Migration flows: construction and description

In this Appendix, we provide elements of context about migration in China, we describe the construction of migration flows from retrospective questions, and we discuss key descriptive statistics.

B.1 Elements of context

An important feature of China's society is the division of the population according to its household registration or *hukou* status. Chinese citizens are classified along two dimensions: their *hukou* type (*hukou xingzhi*)—agricultural (*nongye*) or non-agricultural (*fei nongye*)—and *hukou* location (*hukou suozaidi*). *Hukou* characteristics, which are recorded in the household registration booklet, may not correspond to actual occupation and location.

Since the inception of the reforms in the late 1970s, rules regarding migration within China have been relaxed. Labor mobility remains subject to legal requirements, e.g., being lawfully employed at destination, but the large flows of internal migrants that have characterized China's recent development show that barriers are low in practice, at least for individual migration. Migrants however do not enjoy the same rights as the locally registered population. Whereas an agricultural *hukou* grants access to land, non-agricultural-*hukou* holders enjoy public services in their cities of registration. Access to welfare benefits and public services (e.g., enrollment in local schools, access to health care, urban pension plans, and subsidized housing) is conditional on being officially recorded as a local urban dweller. Migrants thus face a higher cost of living in cities and are supposed to return to their places of registration for basic services such as education and health care or they are charged higher fees (Song, 2014). Labor outcomes are also affected as local governments may issue regulations restricting access to job opportunities or rely on informal guidelines to employers to favor local permanent residents. This is likely to impede family migration, reduce migrant workers' bargaining power, and lock them in a position of "second-class workers" (Demurger et al., 2009).

Despite the rigidity of the *hukou* system and the persistently low rate of *hukou* conversion, reforms have progressively been introduced during the structural transformation of China. Since the 1980s, China has experienced a gradual devolution of power from the central to local governments in terms of *hukou* policy: rules and implementation vary substantially across places and over time. Provincial governments set guidelines, and specific rules are then determined by prefectures, which in practice hold the most power over *hukou* policy (Song, 2014). Two reforms were

introduced in recent years. First, the distinction between agricultural and non-agricultural *hukou* was abolished within local jurisdictions in about one third of Chinese provinces. Albeit an important evolution, this reform does not affect rural-urban migrants who come from other prefectures. Second, *hukou* conversion rules have been gradually loosened. The main channels to change one's *hukou* from agricultural to non-agricultural used to include recruitment by an SOE, receiving college education or joining the army. These conditions have been relaxed since 2000, especially in small cities and towns that attract fewer migrants (Zhang and Tao, 2012). In larger cities, however, conditions for eligibility are tough, so that *hukou* conversion reforms primarily benefit the richest and highly educated (Song, 2014).

The identification strategy described in Section I allows us to deal with the potential endogeneity of migration policy to local factor demand. The shift-share instrument is indeed likely to be orthogonal to such dynamics.

B.2 Data sources and construction of migration flows

Data description In order to measure migration flows, we use the 2005 1% Population Survey, also called "2005 Mini-Census".[7]

The 2005 1% Population Survey constitutes a 1.3% [*sic*] sample of the population selected from 600,000 primary census enumeration districts using a three-stage cluster sampling (Ebenstein and Zhao, 2015). The sampling weights provided by the National Bureau of Statistics (NBS) account for the underlying proportional probability sampling scheme based on the 2004 population registry of the Public Security Bureau. The 2005 Mini-Census was used to test new ways of recording migration and uses the same questionnaire and definitions as the 2010 Census.

A few caveats are in order. First, the sampling frame of the 2005 1% Population Survey contains information on population by registration. High-immigration areas could thus be under-sampled. Comparing the flows for 2005 in the 2005 Mini-Census and 2010 Census, we indeed find a small discrepancy that we attribute to coverage issues. Second, the 2005 Mini-Census offers a set of variables similar to standard censuses, i.e., prefecture-level information on the place of "current residence" and on *hukou* location. However, the 2005 Mini-Census records the timing of *departure*

[7] After the beginning of the reforms and loosening of restrictions on mobility, there was a growing disconnect between census data focusing on *hukou* location and the rising "floating population" of non-locally registered citizens. The 2000 Population Census was the first census to acknowledge this gap and record migrants' places of residence—provided they had been living there for more than 6 months (Ebenstein and Zhao, 2015). In addition to the place of residence (at the prefecture level in our data), *hukou* location (province level) and *hukou* type, the 2000 Population Census contains retrospective information on the place of residence 5 years before the survey (province level) and the reason for departure if residence and registration *hukou* do not coincide.

from a migrant's place of registration rather than of *arrival* at destination. This is not an issue as long as there is no step migration, i.e., if rural dwellers move directly to their final destinations. Third, the data do not record the place of residence at high enough resolution to unambiguously infer whether a migrant is residing in a rural or urban area. Nevertheless, rural-rural migration represents a small share of emigration from rural areas, mostly explained by marriage—which usually gives right to local registration (Fan, 2008).[8] Fourth, we cannot account for migrants who changed their *hukou* location or type. This is quite innocuous given that *hukou* conversion is marginal.

Migration flow construction The retrospective data on migration spells in the Mini-Census allows us to construct yearly migration flows over the period 2000–2005. These flows are directly observed rather than computed as a difference of stocks as common in the migration literature.

We construct annual migration flows between all prefectures of origin and destination by combining information on the current place of residence (the destination), the place of registration (the origin), and the year in which the migrant left the origin. One advantage of working with those data is that they are representative of the whole population, irrespective of their *hukou* status. However, not all migration spells are observed. We describe below (i) which migration spells are directly observed and which spells are omitted, and (ii) how we can infer some of the unobserved spells and adjust the raw migration flows.

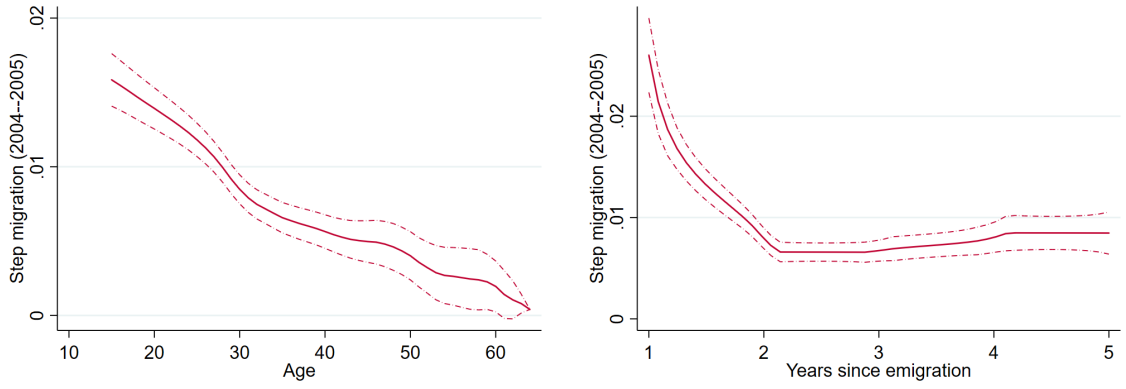
Not all migration spells are observed in the data. We only observe single migration spells, i.e., migration spells in which the interviewed individual is at destination at the time of interview, and whose origin coincides with the *hukou* location. For these individuals, the origin is deduced from their *hukou* location, and the date of their unique relocation is available. All other types of migration histories during the five years preceding the interview are less straightforward to identify. For instance, if one individual were to leave her *hukou* location to city A in 2002 and then relocate to city B in 2005, we would only record the last relocation. In such *step migration* cases, we would correctly specify the departure time from origin, but we would incorrectly attribute arrival dates at destination for the last spell and we would miss arrival in city A. If, instead, one individual were to leave her *hukou* location to city A in 2002 and then return to her *hukou* location by 2005, we would miss her entire migration history. In such *return migration* cases, we would incorrectly omit

[8] In the 2005 Mini-Census, only 6.45% of agricultural-*hukou* holders who migrated between prefectures reported having left their places of registration to live with their spouses after marriage. See Table B2 for further descriptive statistics on reasons for moving.

emigration flows from origins and immigration to destinations.

The incidence of *step migration* and *return migration* spells can, however, be measured. The 2005 Mini-Census records where individuals were living 1 and 5 years before the survey (province level). We can estimate how many migrants report different destinations between 2000 and 2005, which would be a proxy for step migration, and we can observe total return migration between 2000 and 2005, and 2004 and 2005. We first study the importance of step migration. Among migrants who were in their provinces of registration in 2000 and in other provinces in 2005, we compute the fraction that lived in yet another province in 2004. A minority of migrants have changed provinces of destination between 2004 and 2005 (see Figure B1). Step migration is not only low but concentrated in the first year after the first migration spell. In other words, step migration induces errors in arrival and departure dates that are quite small. Adjusting for step migration would require strong assumptions about the intermediate destination, which is not observed in the data; we thus do not correct migration flows for step migration.

Figure B1. Share of step migrants as a function of age and time since departure.



(a) Age. (b) Years since emigration.

Source: 2005 1% Population Survey.
Notes: The sample comprises all working-age (15–64) agricultural-*hukou* holders who were living in a province different from their province of registration in 2004 and left their prefecture of registration less than 6 years prior to the interview.

We then consider the extent of return migration. Among all migrants from rural areas who were living in their provinces of registration in 2000 and in other provinces in 2004, we compute the fraction that had returned to their provinces of registration by 2005. This share is not negligible: In a given year, between 4 and 6% of rural migrants who had left their provinces of registration in the last 6 years go back to their *hukou* locations. Return migration is an important phenomenon, which leads us to underestimate true migration flows and the effect of shocks on emigration.

Because of the retrospective nature of the data, past flows, for instance in 2000 for an individual interviewed in 2005, are mechanically underestimated. In contrast with step migration, it is possible to adjust migration flows and account for return migration. We provide below a description of these adjustments.

Adjusting for return migration requires us to observe the destination and duration-specific yearly rate of return. There is a wide disparity in return rates across destinations. Besides, there are non-negligible compositional adjustments along the duration of the migration spell—as in any survival analysis with censoring. Specifically, the probability for a migrant to return home sharply decreases with the length of the migration spell, mostly reflecting heterogeneity across migrants in their propensity to return. Ignoring such heterogeneity would lead us to underestimate return migration for recent flows and overestimate it for longer spells. To capture variation across destinations and along the length of the migration spell, we make the following assumptions. (i) The "survival" at destination is characterized by a constant Poisson rate f for each migrant. (ii) We suppose that there is a constant distribution of migrant types $H(f)$ *upon arrival*. We allow the distributions to differ across provinces of destination and *hukou* types, i.e., agricultural and non-agricultural. (iii) In order to fit the observed return rates as a function of migration duration, we further assume that:

$$h(f) = \lambda_p^2 f e^{-\lambda_p f}.$$

where λ_p is province- and *hukou* type-specific.

Under the previous assumptions and in a steady-state environment, the evolution of the pool of migrants with duration can easily be computed. In the cross-section (i.e., across all cohorts and not only newly-arrived migrants), the distribution of migrant types is exponential, i.e., $h_c(f) = \lambda_p e^{-\lambda_p f}$, such that the average yearly return rate is $1/\lambda_p$. In all census waves, we observe the *hukou* location, the place of residence five years before the survey, and the place of residence during the survey. This observation allows us to compute the empirical return rate in the cross-section over a period of five years. We calibrate the *hukou*- and province-specific exponential parameter λ_p to match this return rate. Using the calibrated distribution $H(\cdot)$, we can infer the initial flow of migrants from the number of survivors observed k years later and correct for return migration. More precisely, letting $M_{T,k}$ denote the number of migrants arrived in period $t = T - k$ and recorded in period T, the actual number of newly-arrived migrants in $t = T - k$ is $\left[(\lambda_p + k)^2/\lambda_p^2\right] M_{T,k}$.

One concern with this methodology is that we may not precisely capture the duration-dependence in return rates, and thus over- or underestimate return rates for individuals arriving immediately before the interview. We provide an over-

identification test by computing the return probability between 2004 and 2005 for recently-arrived migrants (i.e., between 2000 and 2004), and compare it with the empirical moment. We compute this model-based probability under our baseline specification (B) and under an alternative specification (R) where return rates are assumed to be independent of duration.

Figure B2. Over-identification test for the return migration correction.

Source: 2005 1% Population Survey.
Notes: Blue dots correspond to the baseline specification (duration-dependent return rates). Red dots correspond to an alternative specification, where return rates are assumed independent of migration duration.

Figure B2 displays the model-based return probabilities for recently-arrived migrants against the actual observed return rate. The baseline specification (B, blue dots) matches well the prefecture-level variation in annual return rate for recently-arrived migrants, while the alternative specification (R, red dots) systematically underestimates the incidence of return. Under the alternative specification (R), the return rate after one year is about half the observed rate—a difference due to the fact that the calibration then ignores the difference between the (high) return rate conditional on a short migration spell and the (low) return rate conditional on longer spells. Note that, even under specification (B), there is noise, and some model-based estimates are quite far from the actual return rates. This difference could be due to fluctuations in return rates across years: While the calibration uses the 2000–2005 period, the validation check focuses on 2004–2005 only.

B.3 Migration patterns and the selection of migrants.

Migration patterns over time and across regions Migration patterns vary both over time and across origins and destinations. First, there is a general in-

crease in migrant inflows in 2000–2005, probably related to the decline in mobility costs and the attractiveness of buoyant cities. Migration is mostly rural-urban and long-distance. Over the period, about 80% of the yearly migrant inflows consist of agricultural-*hukou* holders ("rural" migrants), the remainder being urban dwellers originating from other prefectures. About 80% of inter-prefectural rural-urban migrations involve the crossing of a provincial border. The annual inflow of migrants from other prefectures is around 3% of the destination population.

There is a large variation in the spatial distribution of migration inflows and outflows. Some regions (e.g., South Central, which includes the Pearl River Delta) are net recipients and attract a large share of local migrants, while other regions (e.g., North-West) are net senders. As shown in Table B1, there is significant variation in terms of immigration rates across regions, and there is a lot of dispersion of migration spells across destinations.

Table B1. Descriptive statistics of migration flows by region.

	North	North-East	East	South Central	South-West	North-West
Immigration rate (%), 2000						
In prov., out of pref.	0.07	0.20	0.42	0.62	0.40	0.27
In region, out of prov.	0.18	0.08	0.86	1.10	0.27	0.13
Out of region	0.67	0.08	0.74	0.85	0.17	0.43
Immigration rate (%), 2005						
In prov., out of pref.	0.28	0.32	1.22	1.35	0.99	0.53
In region, out of prov.	0.59	0.20	1.78	2.69	0.42	0.21
Out of region	1.70	0.24	2.47	1.90	0.37	0.78
Destination concentration						
HHI, 2000	0.02	0.02	0.03	0.03	0.03	0.02
HHI, 2005	0.03	0.03	0.04	0.04	0.03	0.03

Notes: Migration flows are corrected for return migration. The top and middle panels display yearly migration rates in 2000 and 2005, respectively, by region of destination. Rates are expressed as a share of the total urban population in the region in 2000. The bottom panel provides standardized Herfindahl-Hirschmann Indices (HHI) of destination concentration by region of origin. Prefecture-level HHIs are averaged by region. The index ranges between 0 and 1: an index of 1 indicates that all migrants from a prefecture of origin move to a single prefecture of destination; 0 indicates perfect dispersion.

Selection of migrants We now provide some descriptive statistics on the profile of internal migrants in China relative to non-migrants both in rural and urban areas. Table B2 sheds some light on the motives behind migration. We define migrants as agricultural-*hukou* holders who crossed a prefecture boundary and belong to working-age cohorts (15–64). A vast majority of these migrants (73%) moved away in order to seek work.

Rural-urban migrants are a selected sample of the origin population. We provide some elements of comparison between migrants and stayers in Table B3. Migrants

Table B2. Descriptive statistics from the 2005 Mini-Census.

Reason for moving	Count	Percent of migrants
Work or business	102,388	73.32
Follow relatives	16,454	11.78
Marriage	9,006	6.45
Support from relatives/friends	5,859	4.20
Education and training	1,956	1.40
Other	3,987	2.85

Notes: This table displays descriptive statistics on rural migrants' reasons for migrating.Rural migrants are defined as inter-prefectural migrants with an agricultural *hukou* and aged 15–64.

tend to be younger, more educated, and more often single than the non-migrant rural population. They are also more likely to be self-employed or employees and to work in the private sector. The rural-urban productivity gap appears to be massive, as the migrants' monthly income is more than twice as large as the stayers', which may reflect both selection and different returns to skills in urban and rural areas.

Rural-urban migrants are however also different from urban residents. As is usual with studies of internal migration, we consider in our baseline specification that migrants and locally registered non-agricultural-*hukou* holders are highly substitutable. Table B3 provides summary statistics on key characteristics of inter-prefectural migrants and compares them with the locally registered urban population. Migrants and natives are significantly different on most accounts, the former being on average younger, less experienced, less educated, more likely to be illiterate, and more often employed without a labor contract. Rural-urban migrants are also over-represented in privately owned enterprises and in manufacturing and construction industries: 89% of them are employed in the private sector as against 28% of locally registered *hukou* holders; and the share of rural-urban migrants working in manufacturing and construction is 51% and 9%, as against 20% and 4% for urban residents, respectively. Finally, migrants earn 17% less than urban residents.

To summarize, (i) migrants are selected at origin, (ii) they choose their destination, and (iii) they differ from urban workers along observable characteristics. Our empirical strategy, based on exogenous variation in agricultural prices at origin, is affected by the previous issues as follows. First, compliers are selected and our estimates are a local average treatment effect; we shed some light on this issue in Section C. Second, Chinese rural-urban migrants may not compete with urban residents for the exact same jobs; they may also be less productive for a given job. We further quantify the bias induced by the hypothesis of homogeneous labor productivity in Appendix D.3.

Table B3. Migrant selection (2005 Mini-Census).

	Rural-urban migrants	Local urban *hukou*	Non-migrant rural *hukou*
Age	28.25	37.18	33.81
Female	0.49	0.49	0.50
Married	0.57	0.64	0.56
Education:			
Primary education	0.23	0.17	0.38
Junior high school	0.54	0.30	0.36
Senior high school	0.13	0.26	0.06
Tertiary education	0.02	0.19	0.00
Unemployed	0.02	0.07	0.01
Self-employed/Firm owners	0.13	0.06	0.05
Employees	0.58	0.35	0.08
...of which:			
Public sector	0.11	0.72	0.21
Private sector	0.89	0.28	0.79
Out of the labor force	0.25	0.57	0.45
Monthly income (RMB)	961.0	1155.2	401.4
Hours worked per week	55.08	45.83	45.07
Industry:			
Agriculture	0.05	0.06	0.78
Manufacturing	0.51	0.20	0.08
Construction	0.09	0.04	0.03
Wholesale and retail trade	0.15	0.14	0.04
Other tertiary	0.20	0.51	0.06
Observations	139,813	678,614	1,716,269

Notes: All variables except *Age*, *Monthly income*, and *Hours worked per week* are dummy-coded. The sample is restricted to individuals aged 15–64. Descriptive statistics for *Monthly income (RMB)*, *Hours worked per week*, and industrial sectors are restricted to individuals who reported positive working hours in the past week.

C A shift-share instrument for migration flows

This Appendix develops a stylized, static model of location choice in order to derive our main empirical specification and justify the use of a shift-share instrument for migration inflows to Chinese cities.

C.1 A stylized model

Environment The economy is composed of a unit mass of workers, born at different locations $o \in \Theta$. Let p_o denote the number of workers born at location o.

Workers live for one period only. They are mobile and have heterogeneous preferences over the various locations in which they can provide one unit of labor. For the sake of exposition, we do not explicitly distinguish rural locations from urban ones, and any birth location can be a destination and reciprocally. A worker born at origin o and deciding to work in destination $d \in \Theta$ receives utility:

$$u_{iod} = \frac{b_{iod} w_d}{\kappa_{od}},$$

where b_{iod} is the idiosyncratic amenity draw related to a destination d for worker i, w_d is the exogenous revenue associated with providing labor at destination d, and $\kappa_{od} \geq 1$ is a migration "iceberg" cost capturing migration costs, commuting costs, or cultural differences between origin and destination.

As usual in New Economic Geography framework (see Monte et al., 2018; Bryan and Morten, 2019, for instance), the heterogeneity in preferences for locations across workers is represented by a Frechet distribution. Workers observe their preferences before choosing to work at a certain destination, and the worker-specific idiosyncratic amenity is drawn independently across workers and locations as follows,

$$b_{iod} \sim G_{od}(b) = e^{-B_{od} b^{-\varepsilon}},$$

where B_{od} is a distribution-location parameter which can be interpreted as the relative preferences of the workers from origin o for the specific location d, and $\varepsilon > 0$.

A worker chooses to locate in a destination d if her utility is greater there than in any other locations. With the previous Frechet distribution, the ex-ante probability that a worker chooses destination d is,

$$\mu_{od} = \frac{B_{od} \left(w_d / \kappa_{od}\right)^{\varepsilon}}{\sum_{d \in D} B_{od} \left(w_d / \kappa_{od}\right)^{\varepsilon}}.$$

We assume that $B_{oo} >> B_{od}/\kappa_{od}$ for any destination $d \neq o$; we normalize $kappa_{oo} =$

1. The previous assumption implies that the probability to relocate from an origin o to a different destination is small. Under this assumption, we have:

$$\mu_{od} = \frac{B_{od}\left(w_d/\kappa_{od}\right)^{\varepsilon}}{B_o\left(w_o\right)^{\varepsilon}}.$$

Shocks in revenue and emigration flows We assume that the revenue at any given location is stochastic and follows $w_d = \bar{w}_d\left(1+\hat{w}_d\right)$ where $\bar{w}_d$ is a fixed parameter and $\hat{w}_d$, satisfying $E\left[\hat{w}_d\right] = 0$, is a (small) shock that is known to all workers before they decide to work in a given destination. Letting $\bar{\mu}_{od} = \frac{B_{od}(\bar{w}_d/\kappa_{od})^{\varepsilon}}{B_o(\bar{w}_o)^{\varepsilon}}$, we have, at first-order:

$$\mu_{od} = \bar{\mu}_{od}\left(1 - \varepsilon\hat{w}_o + \varepsilon\hat{w}_d\right).$$

The probability for a worker born at origin o to relocate to any other destination is,

$$n_o = \bar{n}_o\left(1 - \varepsilon\left(\hat{w}_o - \frac{\sum_{d\in\Theta\setminus\{o\}}\bar{\mu}_{od}\hat{w}_d}{\bar{n}_o}\right)\right), \tag{1}$$

where $\bar{n}_o = \sum_{d\in\Theta\setminus\{o\}}\bar{\mu}_{od}$. The parameter $-\varepsilon$ can thus be interpreted as the elasticity of emigration to the revenue at origin relative to the revenue at destination for the average migrant.

Shocks in revenue and immigration flows We now transform the previous emigration rate across possible destinations, μ_{od}, into an immigration rate across origins. The total number of workers arriving at destination d is the sum of all arrivals from any origin $o \neq d$, i.e.,

$$\sum_{o\in\Theta\setminus\{d\}}\mu_{od}p_o.$$

Letting $\lambda_{od} = \frac{\bar{\mu}_{od}p_o}{\sum_{o\in\Theta\setminus\{d\}}\bar{\mu}_{od}p_o}$ and $\bar{m}_d = \frac{\sum_{o\in\Theta\setminus\{d\}}\bar{\mu}_{od}p_o}{p_d}$ denote the share of arrivals from origin o and the migrant share at destination, *absent any shocks*, the immigration rate to the initial population at destination d verifies at first order,

$$m_d = \bar{m}_d\left(1 - \varepsilon\left(\sum_{o\in\Theta\setminus\{d\}}\lambda_{od}\hat{w}_o - \hat{w}_d\right)\right). \tag{2}$$

The parameter $-\varepsilon$ is also the elasticity of immigration to the revenue at the origin of the average migrant, relative to the local revenue.

Push, pull factors and a shift-share instrument We now introduce additional sources of shocks: amenity shocks at different locations, $B_{od} = \bar{B}_{od}\left(1 + \hat{B}_{od}\right)$, and shocks to the relocation costs, $\kappa_{od} = \bar{\kappa}_{od}\left(1 + \hat{\kappa}_{od}\right)$. At first order, the immigration rate to the initial population at destination d verifies,

$$m_d = \bar{m}_d\left(1 + \hat{x}_d^D + \hat{x}_d^O\right), \tag{3}$$

where $\hat{x}_d^D$ is a pull factor and $\hat{x}_d^O$ is a push factor, defined by:

$$\begin{cases} \hat{x}_d^D = \varepsilon\hat{w}_d + \sum_{o\in\Theta\setminus\{d\}} \lambda_{od}\left(\bar{B}_{od} - \varepsilon\hat{\kappa}_{od}\right) \\ \hat{x}_d^O = \sum_{o\in\Theta\setminus\{d\}} \lambda_{od}\left(-\hat{B}_{oo} - \varepsilon\hat{w}_o\right) \end{cases}$$

The immigration rate is driven by a set of pull and push shocks. Immigration is expected to increase following positive shocks to revenue and amenity at destination and negative shocks to relocation costs for the average immigrant. Immigration is expected to increase following negative shocks to amenities and revenues at the origin of the average immigrant.

In practice, most of the variation underlying immigrant flows to cities relates to labor demand at origin and destination, and to barriers to migration (Bryan and Morten, 2019; Tombe and Zhu, 2019), even though a few recent contributions have highlighted the role of amenities, and pollution in particular, in driving labor flows to Chinese cities (Chen et al., 2017; Freeman et al., 2019). The researcher interested in evaluating the causal effect of migration flows on the structure of production, i.e., $y_d = f(m_d)$, faces an identification issue.

To isolate variation in the relocation of workers that is orthogonal to labor demand and consumption patterns in cities, we consider exogenous variation in revenue at origin s_o, and we construct a shift-share instrument, z_d, for the immigration rate m_d by combining origin shocks with the migration patterns,

$$z_d = \sum_{o\in\Theta\setminus\{d\}} \lambda_{od} s_o. \tag{4}$$

The shift-share design transforms shocks at origins in a similar way as the model allocates migration flows, but it requires exogenous shocks pushing migrants toward cities. We discuss the construction of these shocks next.

C.2 Shocks to rural livelihoods

The shift-share design relies on exogenous variation in agricultural livelihoods s_o. This section describes the construction of a shift from variation in cropping patterns combined with innovations in the World demand for commodities.

Cropping patterns across Chinese prefectures Each Chinese prefecture has a distinct agricultural portfolio, resulting from water management, other local technologies, inherent soil characteristics, and climate. To isolate a component in the agricultural portfolio that is orthogonal to adjustments in land use, technological choice or worker-specific productivity, we estimate cropping patterns using a measure of land use at baseline and a measure of potential yield, as calculated from soil and climate characteristics. Harvested areas come from the 2000 World Census of Agriculture, which provides a geo-coded map of harvested areas for each crop at a 30 arc-second resolution (approximately 10 km). We overlay this map with a map of prefectures and construct the total harvested area h_{co} for a given crop c and a given prefecture o. Yields come from the Global Agro-Ecological Zones (GAEZ) Agricultural Suitability and Potential Yields dataset. The GAEZ dataset uses information on crop requirements (i.e., stage-specific crop water requirements) and soil characteristics (i.e., the ability of the soil to retain and supply nutrients) to generate the potential yield for each crop and soil type, under different levels of input and both for rain-fed and irrigated agriculture. We use the high-input scenarios and weight the rain-fed and irrigated yields by the share of rain-fed and irrigated land in harvested areas in 2000 to construct potential yield q_{co} for each crop c and prefecture o.

Table C1 shows the variation in harvested areas across prefectures, by crop and by region. We focus on the four most important crops—rice, wheat, maize, and soy—and on the high-input scenarios. As expected, some crops are more spatially concentrated than others, within the different regions. Rice, for instance, is absent from the colder and drier northern regions. Table C1 however shows that there is substantial regional variation, and no crop is cultivated in a single region, or a region specializing in a single crop. A large part of the cross-sectional variation that we exploit does not come from regional differences, but from more local and granular disparities across prefectures.[9]

International price variation and domestic prices The construction of our shocks to rural livelihoods relies on exogenous variation in World demand for com-

[9] An illustration of these regional differences is also provided in Figure 2 of the paper.

Table C1. Harvested areas across Chinese regions.

	North	North-East	East	South Central	South-West	North-West
Rice, rain-fed	0.000	0.001	0.025	0.038	0.024	0.000
Rice, irrigated	0.129	0.393	0.965	0.725	0.496	0.080
Wheat, rain-fed	0.070	0.018	0.193	0.142	0.155	0.082
Wheat, irrigated	0.652	0.041	0.696	0.780	0.267	0.333
Maize, rain-fed	0.134	0.383	0.218	0.179	0.307	0.093
Maize, irrigated	0.393	0.176	0.318	0.275	0.066	0.163
Soy, rain-fed	0.046	0.097	0.121	0.062	0.092	0.035
Soy, irrigated	0.067	0.021	0.065	0.039	0.015	0.024

Notes: This table displays the between-prefecture variation (measured by the standard deviation and averaged by region) in harvested area for the main crops under irrigated and rain-fed agriculture. Harvested area refers to the normalized area under cultivation.

modities, as captured by producer prices at the farm gate in other countries than China. In our baseline analysis, we isolate an innovation $\hat{\varepsilon}_{ct}$ in the (log) nominal prices, p_{ct}, by running the following AR(1) specification,

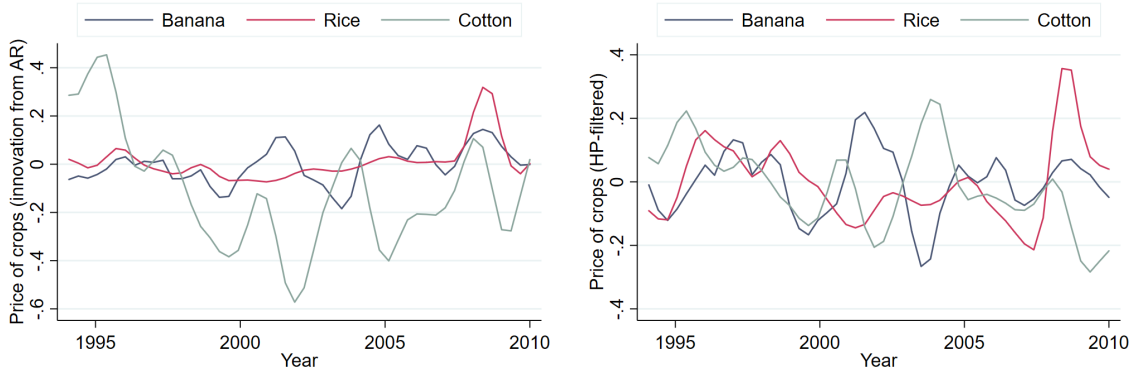
$$p_{ct} = \theta p_{ct-1} + \eta_t + \nu_c + \varepsilon_{ct}$$

We combine these innovations with the previous cropping patterns to construct an exogenous shock to potential agricultural revenue in prefecture o and year t. The shock to potential agricultural revenue between 2000 and 2005 is the average of these yearly shocks. Such strategy hinges on two assumptions.

A first assumption is that short-term fluctuations in international crop prices are quantitatively relevant. Figure C1 plots the evolution of international prices for a selection of crops and shows that there are large swings followed by a gradual return to the mean. Importantly, many different crops display (uncoordinated) fluctuations over time. We interpret these short-term fluctuations as random shocks on the international market due to fluctuations in world demand for each crop.

The second assumption is that local prices are not insulated from world market fluctuations, as captured by fluctuations in producer prices in other countries. The pass-through between international price variations and prices faced by local producers may be mitigated by cooperatives or local policy adjustments. Table C2 displays the correlation between Chinese domestic prices and international prices for different crops in different years. A 10% increase in international prices yields a 2% hike in domestic prices, which constitutes a non-negligible pass-through from the international to domestic markets.

Figure C1. Price deviations from trends on international commodity markets, 1994–2010.



(a) Innovations from an AR(1).

(b) HP-filtered prices.

Source: authors' calculations using the World Bank Commodities Price Data ("The Pink Sheet").
Notes: These series represent the AR(1) and Hodrick-Prescott residuals of the logarithm of international commodity prices, computed separately for three commodities: banana, rice, and cotton.

Table C2. Correlation between international crop prices and local Chinese prices/production.

	Price (producer gate in China)	
	(1)	(2)
Price (international)	0.225	0.182
	(0.026)	(0.042)
Observations	245	245
Year fixed effects	No	Yes

Notes: Standard errors are reported between parentheses and clustered at the crop level. The unit of observation is a crop × year. The crops are banana, cassava, coffee, cotton, barley, groundnut, maize, millet, oats, potato, lentil, rapeseed, rice, sorghum, soybean, sugar beet, sugar cane, sunflower, cabbage, tea and wheat. The sample is an unbalanced panel between 1991 and 2016. Both regressions include crop fixed effects, time trends, the (log) harvested area, and the (log) yield, and are weighted by the value of exports (in tonnes) over the period 1995–2010.

C.3 Sensitivity analysis and robustness checks

This section describes a sensitivity analysis of the shift-share design; we leave the sensitivity of the second stage of the analysis to Appendix F.

The shift The baseline specification isolates an innovation in crop prices from running an AR(1) regression on Agricultural Producer Prices from the FAO, averaged across all producing countries but China. We provide in Table C3 a sensitivity analysis where the price shock is constructed: (i) using commodity prices as extracted from the World Bank Commodities Price Data ("The Pink Sheet", based on prices

on international market places); (ii) restricting the set of crops to the 17 commodities/crops for which the commodity described in the Agricultural Producer Prices data directly matches the one provided by the FAO for harvested area (thereby excluding barley, oats, lentil and cabbage); (iii) using an AR(2) specification instead of an AR(1) in order to isolate a residual in (log) prices; (iv) using a HP filter to isolate a yearly shock to (log) prices. We replicate the results of Table 2 in Panels A and B of Table C3, i.e., we predict emigration with the price shock (the shift) and immigration with the shift-share instrument under these different specifications.

Table C3. Origin-based migration predictions—sensitivity to the construction of the price shock.

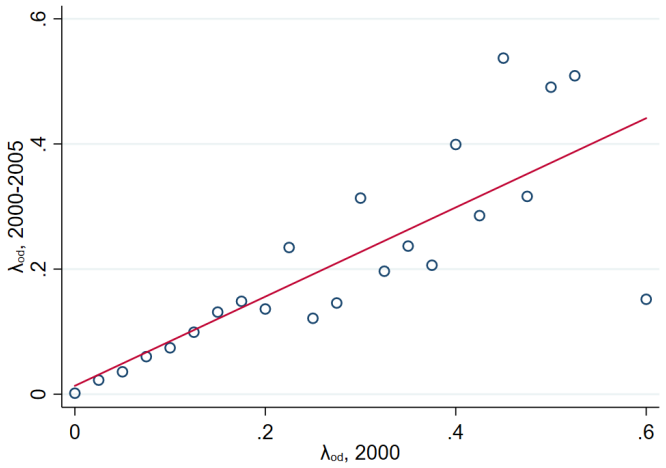
Specification	World Bank	Restricted crops	AR(2)	HP filter
	(1)	(2)	(3)	(4)
Panel A: Predicting emigration				
Price shock	-0.066	-0.081	-0.093	-0.130
	(0.024)	(0.007)	(0.016)	(0.014)
Observations	335	335	335	335
Specification	World Bank	Restricted crops	AR(2)	HP filter
	(1)	(2)	(3)	(4)
Panel B: Predicting immigration				
Shift-share instrument	-1.933	-0.597	-1.307	-1.139
	(0.608)	(0.135)	(0.346)	(0.268)
Observations	315	315	315	315

Notes: Robust standard errors are reported between parentheses. In Panel A, the dependent variable is the number of rural emigrants to urban areas in other prefectures, divided by the number of rural residents at origin. In Panel B, the dependent variable is the number of rural immigrants from other prefectures divided by the number of urban residents at destination. In column 1, the price shock is constructed using World Bank "Pink Sheet" prices; in column 2, the price shock is constructed using a restricted list of crops (i.e., crops which do not require any inference to match harvested commodities and traded commodities); in column 3, the price shock is constructed from isolating an innovation with an AR(2) specification (instead of an AR(1) in the baseline); in column 4, the price shock is constructed using a Hodrick-Prescott filtering (with a parameter of 14,400) on the (log) price of each commodity. See Section I and Equations (2) and (4) for a more comprehensive description of the two specifications.

The share In the baseline specification, we use migration patterns from earlier cohorts, present at destination in 2000, to construct exogenous probabilities to migrate from each origin to each destination. The allocation of earlier migrants across destinations is very predictive of the allocation of migrants in 2000–2005, as shown

in Figure C2. One advantage of using earlier cohorts is that the shares, $\{\lambda_{od}\}$, are independent of amenity or productivity shocks at destination.

Figure C2. Origin-destination migration—the role of previous migration patterns.



Notes: The probabilities for a migrant in d to come from a certain prefecture o, in 2000–2005 (y-axis), and before 2000 (x-axis), are constructed with the 2005 1% Population Survey. Observations are origin × destination couples, and they are grouped by bins of previous migration incidence (bins of 0.025).

One concern with the baseline specification is that these shares may still be correlated with future outcomes through the delayed effects of earlier migration waves (Jaeger et al., 2018). While our identification strategy does not rely on shares being exogenous (as in Goldsmith-Pinkham et al., forthcoming) but on shifts being exogenous (Adão et al., 2019; Borusyak et al., 2018), we provide a robustness check for our first stage in column 1 of Table C4 using the origin of migrants arrived before 1995 and still present at destination in 2000 in order to estimate migration patterns between origins and destinations.

An alternative to this procedure is to estimate a gravity model to predict previous migration (as in Boustan et al., 2010) and use this prediction to redistribute emigration flows across the various destinations. We create a measure of travel cost t_{od} between origin o and destination d using the road and railway networks at baseline. We then predict the migration patterns from earlier cohorts, λ_{od}, using this distance together with a measure of population at destination. This procedure gives us a prediction $\tilde{\lambda}_{od}$ that we can combine with the shifts to generate a shift-share instrument, as in Equation (4). This shift-share instrument also provides a strong first stage for our analysis (see column 2 of Table C4).[10]

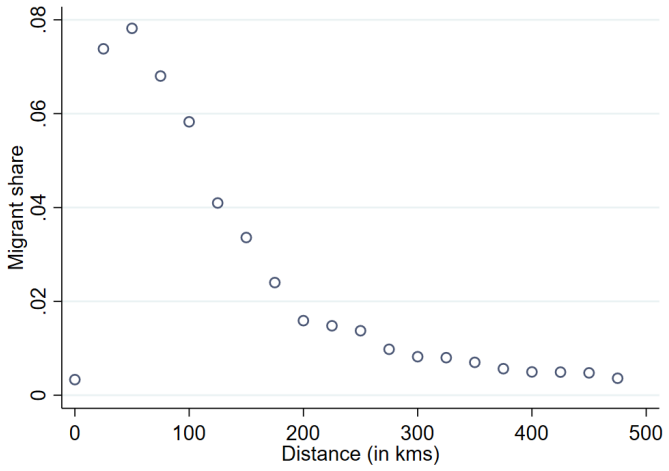
[10]Figure C3 offers visual evidence of the distance gradient in preferred migration routes. There is a strong and significant inverse relationship between the share of migrants from origin o to destination d (among all migrants from o) and distance between o and d.

Table C4. Origin-based migration predictions—sensitivity to the construction of the matrix of migration patterns.

	Pre-1995 migrant shares (1)	Gravity model (2)
Shift-share instrument	-0.864	-0.911
	(0.292)	(0.251)
Observations	315	315

Notes: Robust standard errors are reported between parentheses. The dependent variable is the number of rural immigrants from other prefectures divided by the number of urban residents at destination. In column 1, the matrix of migration patterns is constructed using the stock of migrants at destination in 2000, having arrived in 1995 or before; in column 2, the matrix of migration patterns is constructed from a gravity model. See Section I and Equation (4) for a more comprehensive description of the specification.

Figure C3. Origin-destination migration—the role of distance.



Notes: Migration flows constructed with the 2005 Mini-Census. Observations are origin × destination couples and grouped by bins of distance (25 km).

Other measures of migration flows In the baseline specification, we use migrant flows of workers between 15 and 64 years old and who crossed a prefecture boundary to construct the migration rates and the patterns of settlement. We further rely on migration flows corrected for return migration. In Table C5, we depart from this baseline and allow for various definitions of a migration spell. We show the relationship between the immigration rate and the shift-share instrument when we define migrant flows: (i) irrespective of the migration motive (column 1), (ii) based on males only (column 2), (iii) based on individuals with secondary education or less (column 3), (iv) restricting migration spells to those occurring between

prefectures distant of at least 100 km (column 4); (v) restricting migration spells to those occurring between prefectures distant of at least 300 km (column 5); (vi) using unadjusted measures of migration flows, i.e., raw flows not corrected for return migration (column 6). The relationship between the shift-share instrument and actual migration rates is found to be robust across all specifications.

Table C5. Origin-based migration predictions—sensitivity to the definition of migration spells and migration patterns.

Sample	All	Males	Low-edu.	>100 km	>300 km	Raw
	(1)	(2)	(3)	(4)	(5)	(6)
Shift-share	-1.621	-0.802	-1.363	-1.648	-0.879	-1.379
	(0.425)	(0.220)	(0.375)	(0.461)	(0.479)	(0.378)
Observations	315	315	315	315	315	315

Notes: Robust standard errors are reported between parentheses. The dependent variable is the number of rural immigrants from other prefectures divided by the number of urban residents at destination. In column 1, all migration spells are considered; we restrict the sample to migration spells involving males only in column 2 and low-education individuals only in column 3; we consider migration spells between prefectures distant of at least 100 km in column 4 (resp. 300 km in column 5); we do not adjust the migration flows for return migration in column 6. See Section I and Equation (4) for a more comprehensive description of the specification.

We finally provide evidence that the relationships between the price shock and emigration, and between the shift-share instrument and immigration to cities, are not driven by our treatment of outliers. In the baseline specification, we apply a 99% winsorization to emigration and immigration rates. In Table C6, we replicate Table 2 with the actual, uncensored emigration and immigration rates: both relationships are left unchanged by our censorship procedure.

Migrant selection Throughout the paper, we consider the migrant population as being homogeneous, if not in their preferences (illustrated by origin-specific migration patterns), at least in their labor supply. However, the heterogeneity in preferences for locations may correlate with the heterogeneity in labor supply. For instance, older individuals may be less productive in manual work, and less likely to be tempted by a migration spell to the city. The shift-share instrument may not only affect the size of migrant flows across prefectures but their composition: following a very negative shock to agricultural returns, migrant flows may include "unusual" migrants. We investigate the differential selection of migrants along the agricultural shock in Table C7, in which we regress the average characteristics of migrants between 2000 and 2005 on the price shock. We find that a 10% decrease in agricultural returns—triggering emigration, as shown in Table 2, for instance—increases

Table C6. Origin-based migration predictions—sensitivity to outliers.

	Emigration rate (1)	Immigration rate (2)
Price shock	-0.118	
	(0.019)	
Shift-share instrument		-2.303
		(0.797)
Observations	335	315

Notes: Robust standard errors are reported between parentheses. In column 1, the dependent variable is the number of rural emigrants to urban areas in other prefectures divided by the number of rural residents at origin. Compared to Table 2, migrant flows are not winsorized. In column 2, the dependent variable is the number of rural immigrants from other prefectures divided by the number of urban residents at destination. See Section I and Equations (2) and (4) for a more comprehensive description of the two specifications.

the probability that: (i) migrants are males by 0.5 p.p. (column 1), (ii) migrants have university education by 0.6 p.p. (column 2), (iii) migrants are not married by 1.8 p.p. (column 3). All these effects are however quite small. Column 4 shows that a 10% drop in agricultural returns decreases the age of the average migrant by 0.022 years, an effect indistinguishable from 0. Finally, a 15% drop in agricultural returns increases the probability that the average migrant returns between 2004 and 2005 by about 1 percentage point. The marginal migrant is slightly less attached to destinations than the average migrant.

Table C7. Origin-based migration predictions—selection of migrants.

Characteristics	Male (1)	Low-education (2)	Married (3)	Age (4)	Return rate (5)
Price shock	-0.050	0.059	0.173	0.220	-0.068
	(0.037)	(0.046)	(0.067)	(1.354)	(0.027)
Observations	334	334	334	334	334

Notes: Robust standard errors are reported between parentheses. In columns 1–3, the dependent variable is the share of migrants with the following characteristics: male (column 1), educational attainment lower than high school (column 2), and married (column 3); in column 4, the dependent variable is the age of the migrant in 2005; and in column 5, the outcome is the share of returnees between 2004 and 2005 among inter-provincial rural migrants who emigrated between 2000 and 2004 (the price shock is then defined over the period 2000–2004).

D Migration and factor productivity

This Appendix develops a quantitative framework of firm production, in which there are sector-specific complementarities between capital and labor (Oberfield and Raval, 2014), in order to estimate the impact of migration on factor productivity.

The section is organized as follows. We first derive equations characterizing the optimization program of firms, and we describe the steps for estimating the model parameters: the elasticity of substitution between capital and labor; factor shares; and the elasticity of substitution between product varieties. Second, we describe the effect of migration on factor productivity at destination. Third, we discuss the possible biases related to the imperfect observation of the firm technology and of labor efficiency.

D.1 Quantitative framework

Environment The economy is composed of D prefectures. In each prefecture d, the economy is divided into sectors within which there is monopolistic competition between a large number of firms. The final good is produced from the combination of sectoral outputs, and each sectoral output is itself a CES aggregate of firm-specific differentiated goods. Firms face iso-elastic demand with σ denoting the elasticity of substitution between the different varieties of the sectoral good. In what follows, we drop prefecture indices for the sake of exposition.

Total sectoral output in a product market (sector $\times$ prefecture) is given by the following CES production function:

$$y = \left[\sum_i x_i y_i^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}, \tag{1}$$

where x_i captures consumer preferences for product variety i. Each firm i thus faces the following demand for its product variety i:

$$y_i = (p_i/p)^{-\sigma} x_i^{\sigma} y \tag{2}$$

where p_i is the unit price for variety i, and p is the price index at the product market level. We assume that a firm i produces according to a CES production function:

$$y_i = A_i \left[\alpha k_i^{\rho} + (1-\alpha) l_i^{\rho} \right]^{\frac{1}{\rho}}, \tag{3}$$

where ρ, governing the elasticity of substitution between capital and labor, is as-

sumed constant over time and within sector, and α characterizes the sector-specific share of capital in production.

Firm i maximizes the following program,

$$\pi = \max_{p_i, y_i, l_i, k_i} \{p_i y_i - w l_i - r k_i\}, \tag{4}$$

subject to demand for its specific variety (2) and the production function (3). The previous program determines factor demand across sectors,

$$\begin{cases} (1-1/\sigma)\dfrac{\alpha k_i^{\rho}}{\alpha k_i^{\rho} + (1-\alpha) l_i^{\rho}} p_i y_i = r k_i \\ (1-1/\sigma)\dfrac{(1-\alpha) l_i^{\rho}}{\alpha k_i^{\rho} + (1-\alpha) l_i^{\rho}} p_i y_i = w l_i. \end{cases}$$

Aggregating at the sector level and using a first order approximation brings:

$$\begin{cases} (1-1/\sigma)\dfrac{\alpha \overline{K}^{\rho}}{\alpha \overline{K}^{\rho} + (1-\alpha) \overline{L}^{\rho}} \overline{PY} = \overline{rK} \\ (1-1/\sigma)\dfrac{\beta \overline{L}^{\rho}}{\alpha \overline{K}^{\rho} + (1-\alpha) \overline{L}^{\rho}} \overline{PY} = \overline{wL}, \end{cases}$$

which characterize factor demand at the sector level.

Estimation The following fundamentals of the model need to be estimated: the degree of substitution between capital and labor (ρ), the factor intensity (α), the elasticity of substitution between product varieties (σ).

The key parameter is the elasticity of substitution between factors: the factor intensity and the elasticity of substitution between product varieties can be imputed from factor shares and the ratio of profits to revenues. Indeed, we can infer within-product competition σ from the mere observation of aggregate profits and aggregate revenues:

$$1/\sigma = \overline{\Pi}/\overline{PY}.$$

In a similar fashion, we can use the aggregate first-order condition relating labor costs to revenues in order to identify the capital share α, once ρ is known:

$$\alpha = \frac{(1-X)\overline{L}^{\rho}}{(1-X)\overline{L}^{\rho} + X\overline{K}^{\rho}},$$

where $X = \overline{wL} / \left[(1-1/\sigma)\overline{PY}\right]$.

The issue is to provide an estimate for the elasticity of substitution. One option

is to use estimates provided by Oberfield and Raval (2014) for the United States in 1997. Another option is to use their estimation strategy on our dataset. To do so, we combine the two first-order conditions and derive the firm-specific relative factor demand:

$$\ln\left(k_i/l_i\right) = \frac{1}{1-\rho}\ln\left(\frac{\alpha}{1-\alpha}\right) + \frac{1}{1-\rho}\ln\left(w/r\right) + \varepsilon_i,$$

where ε_i is a noise which captures unobserved firm heterogeneity (e.g., firm-specific relative factor intensity as described in Section D.3), which we assume, as in Oberfield and Raval (2014), to be normally distributed within a sector and a prefecture.

Identifying the elasticity of substitution from the previous relationship is challenging because omitted variation (e.g., a labor productivity shock) may influence relative factor prices and relative factor use. However, the arrival of migrants shifts the relative price of labor downward, an effect that is orthogonal to omitted variation related to labor demand. We can identify the parameter ρ using the variation in relative factor prices across prefectures induced by the shift-share instrument z_{dt}. More specifically, the strategy for estimating the elasticity of substitution relies on the relative factor demand equation,

$$\ln\left(k_{idt}/l_{idt}\right) = \frac{1}{1-\rho}\ln\left(\frac{\alpha}{1-\alpha}\right) + \frac{1}{1-\rho}\left(w_{idt}\right) + \eta_t + \nu_d + \varepsilon_{idt}, \tag{5}$$

where i denotes a manufacturing establishment, d the prefecture and t the year, and w_{idt}, the average compensation rate in establishment i at time t, is instrumented by the shift-share instrument z_{dt}. The identification of Equation (5) hinges on variation across prefectures and over time in relative factor prices and requires the following assumptions. First, we assume that ρ and α are constant over time and across all firms. Second, the rental cost of capital is not observed and possible differences in access to capital across prefectures are assumed orthogonal to the shift-share instrument. We report the estimated ρ in Table D1.

D.2 Effect of migration on factor productivity

The previous estimates for sectoral production allow us to construct factor productivity from the observation of output and factor use at the establishment level. In this section, we use model-based measures of factor productivity to estimate the impact of immigration on productivity at destination. We estimate Equation (5) using the marginal revenue product of labor, marginal revenue product of capital and total factor productivity in revenue terms as dependent variables (all in logs).

In Panel A of Table D2, we use factor productivity as constructed with sectoral

Table D1. Elasticity of relative factor use to the relative factor price.

	Relative factor use
Labor cost	0.810
	(0.246)
Observations	191,316
F-stat. (first stage)	5.49

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The sample is composed of the 31,886 firms present every year in the NBS firm census between 2001 and 2006. All specifications include year fixed effects. *Labor cost* is the average compensation rate in the establishment—$\ln(w_{idt})$ in Equation (5),—and *Relative factor use* is $\ln(k_{idt}/l_{idt})$.

Table D2. Impact of migration inflows on product of factors.

	Labor pr. (1)	Capital pr. (2)	Total fact. pr. (3)
Panel A: Oberfield and Raval (2014) for US 1997			
Migration	-0.748	0.138	-0.231
	(0.172)	(0.087)	(0.076)
Observations	30,556	30,556	30,556
	Labor pr. (1)	Capital pr. (2)	Total fact. pr. (3)
Panel B: Own estimate			
Migration	-1.041	0.364	-0.253
	(0.234)	(0.124)	(0.077)
Observations	30,556	30,556	30,556

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. Each cell is the outcome of a separate regression. The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006 and for which we could estimate the elasticity of substitution. *Labor pr.* is the (log) marginal revenue product of labor; *Capital pr.* is the (log) marginal revenue product of capital; *Total fact. prod.* is the (log) total factor productivity in revenue terms. These quantities are computed using a CES production function with the elasticities of substitution of Oberfield and Raval (2014) (United States, 1997, see Panel A) and our own estimate (Panel B).

estimates of ρ from Oberfield and Raval (2014) for the United States in 1997. The first column reports how the marginal return to labor responds to migrant inflows. The elasticity with respect to migration is about -0.75. In parallel, the marginal revenue product of capital responds positively to the labor supply shift (column 2). There is some evidence of a negative effect on total factor productivity (column 3). In

Panel B of Table D2, we use our own estimates of ρ to construct factor productivity, and find qualitatively similar results as in Panel A. Capital and labor are more complementary than what Oberfield and Raval's (2014) estimates would imply; the arrival of immigrants without further capitalization affects labor productivity more strongly than in the previous specification.

This exercise may however suffer from two biases that we describe next: firm technology may be endogenous and result from a choice (as illustrated by the re-optimization of production lines shown in Section III); labor may not be homogeneous.

D.3 Heterogeneous firm technology and labor efficiency

This section discusses the possible biases induced by heterogeneous firm technology, and by heterogeneity in labor efficiency between migrants and urban residents.

Technology choice We consider the previous framework and assume that individual firms are characterized by (residual) technological choices—different product varieties require different production technologies.[11] Within a sector, some establishments rely on a labor-intensive technology and are labor-abundant, while others are capital-abundant.

$$y_i = A_i \left[\alpha_i k_i^{\rho} + \beta_i l_i^{\rho}\right]^{\frac{1}{\rho}},$$

where ρ, governing the elasticity of substitution between capital and labor, is assumed constant over time and within sector, and (α_i, β_i) characterizes the firm-specific technology—unobserved to the econometrician. We rationalize differences in factor use across production units by technological choices: individual firms produce different varieties, involving more or less labor-intensive production lines.

Heterogeneous firm technology has two implications on the estimation of our production estimates. First, it rationalizes the heterogeneity in factor demand across manufacturing establishments within the same market. Indeed, for a given technology (α_i, β_i), firm i maximizes the following program,

$$\pi\left(\alpha_i, \beta_i\right) = \max_{p_i, y_i, l_i, k_i} \left\{p_i y_i - w l_i - r k_i\right\}, \tag{6}$$

subject to demand for its specific variety and the production function. This maxi-

[11] This feature may capture the wide dispersion in relative factor use within prefecture and industry—See Appendix A.

mization program leads to the following firm-specific relative factor demand:

$$\ln\left(k_i/l_i\right) = \frac{1}{1-\rho}\ln\left(\frac{\alpha}{1-\alpha}\right) + \frac{1}{1-\rho}\ln\left(w/r\right) + \frac{1}{1-\rho}\ln\left(\frac{\alpha_i(1-\alpha)}{\alpha\beta_i}\right),$$

where the last term can be interpreted as the noise in Equation (5).

Second, firms may select their product varieties as a response to changes in factor costs. They would then maximize the indirect profit, $\pi\left(\alpha_i, \beta_i\right)$, subject to a sector-specific technological frontier,

$$\left[\left(\alpha_i/\alpha\right)^{\tau} + \left(\beta_i/(1-\alpha)\right)^{\tau}\right]^{\frac{1}{\tau}} \leq 1, \tag{7}$$

where τ is the curvature of the technological frontier. With endogenous technology choice, optimal factor demand verifies:

$$\ln\left(\frac{k_i}{l_i}\right) = -\frac{\tau}{\left(\tau-1\right)\left(1-\rho\right)-\rho}\ln\left(\frac{\alpha}{1-\alpha}\right) - \frac{1}{1-\rho-\frac{\rho}{\tau-1}}\ln\left(\frac{r}{w}\right).$$

This equation implies that, if the technological frontier is concave ($\tau > 1$), the elasticity of factor demand is larger in absolute value than with a fixed technology. At heart, following an outward shift in labor supply, firms do not only substitute labor for capital until they adjust marginal product of factors as evaluated at their current technology; they also eventually adjust their technology toward more labor-intensive product varieties and this effect adds to the direct impact.

This adjustment may explain why the wedge between the marginal product of labor and its marginal cost decreases with immigrant inflows, and why capital productivity slightly increases. Firms become more labor-abundant in prefectures experiencing large migrant inflows, specifically so by adopting more labor-intensive production lines. A competing explanation is that we assume away any productivity difference between migrant and resident workers in the baseline; any discrepancy between the productivity of urban residents and rural-urban migrants would generate a bias in the estimated effect of migrant inflows on factor productivity.

Heterogeneous labor and the impact of migration In the baseline model of production, labor and wage rates are measured in efficient units. In the data, however, the corresponding variables (*employment* and *labor cost*) do not allow us to distinguish between worker types, and we cannot compute efficient labor units. This limitation may bias the estimated effect of migrant inflows on factor productivity. More specifically, the decrease in the observed labor productivity may reflect the lower productivity of the marginal migrant.

In this extension, we allow workers to differ in productivity and assume that these differences are observable to the manufacturing firm. Consider two worker types, residents indexed by r and migrants indexed by m, and let $h = l_r + \gamma l_m$ denote efficient labor units, where $\gamma < 1$ and $l = l_r + l_m$ is observed employment. For the sake of exposition, we omit the indices and consider the production technology,

$$y = A\left[\alpha k^{\rho} + (1-\alpha)h^{\rho}\right]^{\frac{1}{\rho}}.$$

The first-order conditions give us:

$$\begin{cases} MPL = (1-1/\sigma)\dfrac{\alpha k^{\rho-1}}{\alpha k^{\rho} + (1-\alpha)h^{\rho}} py = r \\ MPK = (1-1/\sigma)\dfrac{(1-\alpha)h^{\rho-1}}{\alpha k^{\rho} + (1-\alpha)h^{\rho}} py = w, \end{cases}$$

where $w = w_r = w_m/\gamma$ is the wage rate.

In the empirical exercise, we use the observed revenues py, the total employment cost wh, the observed capital k, and the observed units of labor l in order to compute the labor cost,$\widehat{w} = w\left(\dfrac{h}{l}\right)$, returns to factors,

$$\widehat{MPL} = (1-1/\sigma)\frac{\alpha k^{\rho-1}}{\alpha k^{\rho} + (1-\alpha)l^{\rho}} py = MPL\left(\frac{l}{h}\right)^{\rho-1}\frac{\alpha k^{\rho} + (1-\alpha)h^{\rho}}{\alpha k^{\rho} + (1-\alpha)l^{\rho}}$$

$$\widehat{MPK} = (1-1/\sigma)\frac{(1-\alpha)l^{\rho-1}}{\alpha k^{\rho} + (1-\alpha)l^{\rho}} py = MPK\frac{\alpha k^{\rho} + (1-\alpha)h^{\rho}}{\alpha k^{\rho} + (1-\alpha)l^{\rho}},$$

and revenue-based total factor productivity, $\widehat{pA} = pA\left(\dfrac{\alpha k^{\rho} + (1-\alpha)h^{\rho}}{\alpha k^{\rho} + (1-\alpha)l^{\rho}}\right)^{1/\rho}$, which all differ from their actual values.

In what follows, we quantify the bias induced by differences in the estimation of the elasticities of these quantities to a marginal increase of the number of migrant workers l_m. For simplicity, we will keep the other factors k and l_r constant. These elasticities are:

$$\frac{\partial \ln(\widehat{w})}{\partial l_m} = \frac{\partial \ln(w)}{\partial l_m} - \frac{(1-\gamma)l_r}{(l_r + \gamma l_m)(l_r + l_m)}$$

for the labor cost,

$$\frac{\partial \ln(\widehat{MPL})}{\partial l_m} = \frac{\partial \ln(MPL)}{\partial l_m} + \frac{\partial}{\partial l_m}\ln\left[\frac{\alpha k^{\rho} + \beta h^{\rho}}{\alpha k^{\rho} + \beta l^{\rho}}\right] + (\rho - 1)\frac{(1-\gamma)l_r}{(l_r + \gamma l_m)(l_r + l_m)}$$

$$\frac{\partial \ln(\widehat{MPK})}{\partial l_m} = \frac{\partial \ln(MPK)}{\partial l_m} + \frac{\partial}{\partial l_m}\ln\left[\frac{\alpha k^{\rho} + \beta h^{\rho}}{\alpha k^{\rho} + \beta l^{\rho}}\right]$$

for the returns to factors, and

$$\frac{\partial \ln(\widehat{pA})}{\partial l_m} = \frac{\partial \ln(pA)}{\partial l_m} + \frac{1}{\rho}\frac{\partial}{\partial l_m}\ln\left[\frac{\alpha k^\rho + \beta h^\rho}{\alpha k^\rho + \beta l^\rho}\right]$$

for the revenue-based total factor productivity. Under the hypothesis that $l_m << l_r$, which induces that our estimate will be an upper bound for the bias, and following a small increase of $\Delta l_m = 1\% l_r$, we have:

$$\begin{cases} \Delta \ln(\widehat{w}) = \Delta \ln(w) - (1-\gamma)\% \\ \Delta \ln \widehat{MPL} = \Delta \ln(MPL) - (1-\gamma)\rho\dfrac{\beta l^\rho}{\alpha k^\rho + \beta l^\rho}\% + (\rho - 1)(1-\gamma)\% \\ \Delta \ln \widehat{MPK} = \Delta \ln(MPK) - (1-\gamma)\rho\dfrac{\beta l^\rho}{\alpha k^\rho + \beta l^\rho}\% \\ \Delta \ln \widehat{pA} = \Delta \ln(pA) - (1-\gamma)\dfrac{\beta l^\rho}{\alpha k^\rho + \beta l^\rho}\%. \end{cases}$$

In order to quantify the bias for the different elasticities, we need to calibrate some parameters. First, the value of $\gamma < 1$ can be retrieved by regressing the (log) wages of all individuals present in the 2005 Mini-Census on a dummy for newly-arrived migrants and a large set of controls, including occupation fixed effects, destination fixed effects, age, education, and gender. This exercise yields $\gamma = 0.80$. Second, the ratio $\beta l^\rho/(\alpha k^\rho + \beta l^\rho)$ is approximately equal to the share of total labor costs over total factor costs, which in China is around 60%. Third, the value of ρ depends on the industry, but for most industries this value ranges between -0.3 and -0.8, and we will use an estimate of -0.5. These calibrated values lead to the following order of magnitude for the (maximum) biases:

$$\begin{cases} \Delta \ln(\widehat{w}) \approx \Delta \ln(w) - 0.20\% \\ \Delta \ln \widehat{MPL} \approx \Delta \ln(MPL) - 0.24\% \\ \Delta \ln \widehat{MPK} \approx \Delta \ln(MPK) + 0.06\% \\ \Delta \ln \widehat{pA} \approx \Delta \ln(pA) - 0.12\%. \end{cases}$$

For an employment effect between 0.3 and 0.4, the elasticities of the labor cost, the returns to labor and capital, and the total factor productivity would need to be corrected at most by -0.07, -0.08, +0.02, and -0.04.

E Additional results on heterogeneity and firm entry/exit

This Appendix provides complementary results to the baseline analysis: on treatment heterogeneity along a set of important firm characteristics (e.g., factor use at baseline, factor product at baseline, firm ownership); on firm entry into, and exit from, the survey of manufacturing establishments; on the wage of urban residents; on the restructuring of production lines.

E.1 Treatment heterogeneity across urban establishments

The analysis of Section II focuses on the change within the average establishment. The shift toward a more labor-intensive production structure may also involve a reallocation of resources across establishments (Dustmann and Glitz, 2015). We now provide evidence on the heterogeneous absorption of migrants in the urban economy and its (limited) aggregate implications.

We study the heterogeneous response in factor demand by interacting migrant inflows with the following firm characteristics at baseline: (a) we label as capital-abundant all firms with a capital-to-labor ratio at baseline above the median in their sector and prefecture; (b) we label as labor-productive all firms with a value added per worker at baseline above the median in their sector and prefecture; (c) we isolate firms within a sector with high degree of substitution between capital and labor (ρ); (d) we isolate firms with educational requirement above the sector median, as calculated from the proportion of workers with high-school attainment or less in 2004; (e) we separate public establishments from privately-owned firms, (f) we isolate establishments with some exports at baseline, in 2000.

We report the heterogeneous effects of migration on our baseline outcomes, and along the previous characteristics, in Table E1. The reduction in labor cost is remarkably homogeneous across firms; all firms seem to face similar labor market conditions. In response to the labor supply shift, we find that capital-abundant firms recruit more than the average firm (column 2). However, firms with higher average labor productivity are less likely to expand in response to the migration shock. Migrant workers are predominantly recruited by "capital-rich" firms in a given sector and location; they are hired by firms where labor productivity is slightly lower.[12]

[12]Interpreting these findings requires understanding the nature of heterogeneity within our survey of manufacturing establishments. First, disparities in labor productivity may arise from unobserved technological differences across production units. In a similar vein, factor use at baseline may reflect firm-specific factor complementarities in production or complementarities in production with other unobserved factors (e.g., skilled labor) that are heterogeneously allocated across establishments. Second, the initial dispersion in productivity across firms may reflect factor market imperfections (as in Hsieh and Klenow, 2009). Firms with a high return to labor may be constrained

This observation contrasts with empirical regularities of firm growth in developed economies: employment flows are typically directed toward productive firms (see Davis and Haltiwanger, 1998, for evidence in the U.S.). A possible explanation is that we study large labor supply shocks, which may have different allocative properties from the smaller idiosyncratic labor demand shocks that usually drive employment growth. Our findings are also different from Dustmann and Glitz (2015), who find that more labor-abundant firms expand relative to capital-abundant firms.

We explore in Panels C and D whether sectoral differences in production matter, notably through the complementarity between labor and capital, or through skill requirements. We do not find that migrant workers sort themselves into sectors with high elasticity of substitution between capital and labor, or with low education requirements. The interaction coefficients are quite small in both cases. We finally interact the immigration rate with a dummy for public firms (Panel E) and exporting firms (Panel F). We do not find that migrants are less likely to be hired in public establishments, in which insiders are likely to receive substantial benefits, and in exporting firms which can arguably expand their production, facing more gradual decreasing returns to scale.

E.2 Aggregation and entry/exit

Aggregation at the prefecture level In Section II, we explore the aggregate effects of factor reallocation across firms by aggregating outcomes at the prefecture $\times$ sector level. Conceptually, this neutralizes the effect of a possible reallocation of factors within sectors. We now report estimates from specification (5) with outcomes aggregated at the prefecture level (see Table E2). The results are similar in magnitude to the previous aggregation at the prefecture $\times$ sector level: the effect of a reallocation of factors across industries on aggregate productivity is negligible, as most of the reallocation occurs within sectors. This finding is consistent with the literature in developed countries (Lewis, 2011; Dustmann and Glitz, 2015).

Entry/Exit In Section II, we present the effect of immigrant flows on aggregate outcomes including firms of the unbalanced sample, and we find that allowing for firm entry and exit magnifies the negative effect of migration on relative factor use

in hiring labor, e.g., because of information asymmetry between job seekers and employers (Abebe et al., 2016; Alfonsi et al., 2017), the intervention of intermediaries, and the prevalence of migrant networks (Munshi, 2003). Capital productivity dispersion may be indicative of capital market distortions (Buera et al., 2011; Midrigan and Xu, 2014). Third, productivity differences across firms may capture entrepreneur characteristics, management practices (Bloom et al., 2013) or differences in the organization of production (Akcigit et al., 2016; Boehm and Oberfield, 2018). Better entrepreneurs or organizations would be captured by high productivity within a sector.

Table E1. Impact of migration inflows on urban firms—heterogeneity across firms.

	Labor cost (1)	Employment (2)	K/L ratio (3)	Y/L ratio (4)
Panel A: Capital to labor ratio				
Migration	-0.124	0.269	-0.363	-0.412
	(0.071)	(0.053)	(0.095)	(0.100)
Migration × High K/L	-0.070	0.078	-0.227	-0.087
	(0.083)	(0.055)	(0.090)	(0.078)
F-stat. (first stage)†	11.97	11.97	11.97	11.97
Panel B: Output to labor ratio				
Migration	-0.116	0.306	-0.363	-0.304
	(0.075)	(0.050)	(0.106)	(0.110)
Migration × High Y/L	-0.068	-0.048	-0.164	-0.308
	(0.088)	(0.065)	(0.091)	(0.086)
F-stat. (first stage)†	11.77	11.77	11.77	11.77
Panel C: Elasticity of substitution				
Migration	-0.146	0.278	-0.457	-0.442
	(0.060)	(0.051)	(0.097)	(0.110)
Migration × High ρ	-0.003	0.040	0.067	0.011
	(0.071)	(0.059)	(0.070)	(0.069)
F-stat. (first stage)†	11.80	11.80	11.80	11.80
Panel D: Educational requirement				
Migration	-0.133	0.268	-0.468	-0.427
	(0.058)	(0.056)	(0.111)	(0.115)
Migration × High share of educated	-0.028	0.058	0.080	-0.021
	(0.065)	(0.044)	(0.076)	(0.075)
F-stat. (first stage)†	11.51	11.51	11.51	11.51
Panel E: Public ownership				
Migration	-0.106	0.233	-0.464	-0.362
	(0.065)	(0.051)	(0.104)	(0.097)
Migration × Public	-0.049	0.006	0.228	-0.225
	(0.132)	(0.117)	(0.182)	(0.169)
F-stat. (first stage)†	8.88	8.88	8.88	8.88
Panel F: Exporting at baseline				
Migration	-0.072	0.357	-0.505	-0.517
	(0.128)	(0.084)	(0.132)	(0.195)
Migration × Exporting	-0.060	-0.100	0.117	0.192
	(0.122)	(0.088)	(0.120)	(0.159)
F-stat. (first stage)†	11.76	11.76	11.76	11.76
Observations	31,886	31,886	31,886	31,886

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. † The IV specification uses two endogenous variables and two instruments; the critical value for weak instruments is then 7.03 (at 10%).

Table E2. Impact of migration inflows on urban firms—sensitivity analysis with aggregate variables at the prefecture level.

	Labor cost (1)	Employment (2)	K/L ratio (3)	Y/L ratio (4)
Panel A: Balanced sample of firms				
Migration	-0.113	0.284	-0.459	-0.475
	(0.062)	(0.071)	(0.125)	(0.120)
Observations	315	315	315	315
F-stat (first)	23.52	23.52	23.52	23.52
Panel B: Unbalanced sample of firms				
Migration	-0.184	0.627	-0.626	-0.626
	(0.083)	(0.159)	(0.120)	(0.184)
Observations	315	315	315	315
F-stat. (first)	21.12	21.12	21.12	21.12

Notes: Robust standard errors are reported between parentheses. The unit of observation is a prefecture. In Panel A (resp. Panel B), the sample is composed of the firms present every year in the NBS firm census between 2000 and 2006 (resp. all firms present in the NBS firm census between 2000 and 2006); outcomes are then aggregated at the prefecture level. *Migration* is the immigration rate, i.e., the migration flow divided by destination population at baseline. *Labor cost* is the (log) compensation per worker including social security. *Employment* is the (log) number of workers within the firm. *K/L ratio* is the (log) ratio of fixed assets to employment. *Y/L ratio* is the (log) ratio of value added to employment.

and labor productivity. This extension provides additional insights into the selection of establishments into survival, or entry.

We first estimate the effect of migration on profits. We consider the balanced sample of firms and construct the ratio of profits to revenues (profitability) and a dummy equal to one if profits are positive. Columns 1 and 2 of Table E3 present the estimates of specification (5) for these two outcomes. The arrival of low-skill workers does affect profitability in the average establishment (see column 1); it also markedly increases the probability that an establishment reports net profits (see column 2), thus mostly benefiting low-profitability establishments. A ten percentage point increase in the immigration rate increases the probability that firms make profit by 1.1 percent. Cheaper labor makes the least profitable firms break even.

We now estimate the more direct effect of migration on firm entry into, and exit from, the survey of manufacturing establishments. Since we only observe firms above a given sales threshold (see Appendix Section A and Figure 1), we only measure entry and exit into and from our sample, which is a combination of actual entry and

exit, and of firms growing into and shrinking out of the sample. We use an additional piece of information, i.e., the year the firm was founded, and check whether the year in which establishments enter the sample corresponds to their first year of operation in order to define entry. Columns 3 and 4 of Table E3 presents the estimates of a specification collapsed at the prefecture × sector level. There is a negative and significant effect on the exit rate from the sample (column 3). This finding is consistent with the previous finding that migration benefits low-profitability firms and increases the probability to declare some positive profits. This profitability effect allows such establishments to survive, or to remain large enough and appear in our sample of above-scale firms. We also find that migration has a negative effect on the probability that a firm is created and appears in the sample (column 4).

Table E3. Impact of migration inflows on urban firms—profitability and entry/exit.

	Profitability (1)	Any profit (2)	Exit rate (3)	Entry rate (4)
Migration	0.043	0.105	-0.119	-0.058
	(0.008)	(0.034)	(0.032)	(0.033)
Observations	31,864	31,886	6,742	6,457
Mean outcome	-	-	.602	.511

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. In columns 1 and 2, the unit of observation is a firm and the sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. In columns 3 and 4, the unit of observation is a sector × prefecture and the sample is composed of all firms present in the firm census at any point between 2000 and 2006. *Migration* is the immigration rate, i.e., the migration flow divided by destination population at baseline. *Profitability* is the change in the ratio of profits to revenues in 2000 and 2006. Profitability is missing for two firms that reported zero revenue. *Any Profit* is the difference between indicator variables equal to 1 if the firm made positive profits in 2000 and in 2006. *Exit Rate* is the share of firms present in 2000 that were no longer in the sample in 2006. *Entry Rate* is the share of firms present in 2006 that were not present in 2000 and were founded between 2000 and 2006. It is missing for sectors × prefectures that were present in 2000 but did not have any more firms in 2006. See Section I and Equation (5) for a description of the IV specification.

E.3 Worker heterogeneity and compositional effects at destination

In the baseline analysis, we interpret the decrease in labor cost as a decline in the equilibrium wage. However, compensation per worker may fall due to changes in the composition of the workforce, as less skilled workers enter the manufacturing sector and potentially displace skilled resident workers (Card, 2001; Monras, 2015). To better capture the migration effect on the labor market outcomes of *residents*,

we use the Urban Household Survey between 2002 and 2006 (see a description in Appendix A) and we estimate:

$$\Delta y_d = \alpha + \beta m_d + \mathbf{X}_d \gamma + \varepsilon_d, \tag{1}$$

where y_d is the average labor market outcome of individuals surveyed in prefecture d between 2002 and 2006, m_d is the immigration rate between 2001 and 2005—instrumented by the shift-share instrument z_d, and $\mathbf{X}_d$ is a vector of average individual characteristics, including marital status, gender, and age.

Panel A of Table E4 presents the results computed from all urban residents. In column 1, the dependent variable is a measure of real hourly wages. The wage of residents falls by 3.0% when the migration rate increases by ten percentage points—an effect that is imprecisely estimated due to the lower number of prefectures in the UHS. In columns 2 to 4, we analyze the possible displacement of urban residents. Rural-urban migration has no significant effect on the allocation of urban residents between wage employment, unemployment, and self-employment, which implies that the urban residents mostly adjust to an immigration shock by accepting lower wages. In Panel B, we derive the same results, but computed from low-skilled residents only, and find qualitatively similar results.

There are various reasons for which the wage of residents would be expected to be *less elastic* to the arrival of immigrants than the average manufacturing wage: labor markets may be partly segmented; the wages of incumbent workers may be more rigid than hiring wages; migrants may be less productive than residents, and the recruitment of lower-productivity workers could account for part of the decline in average labor cost. We find however that the decrease in wages of low-skill residents is larger than the labor cost response estimated using firm-level data (see Table 3). This difference could arise from the migration effect on firms outside our sample of manufacturing establishments, which are formal production units and in which there may exist bargaining frictions inducing (some) nominal downward rigidities.

We also use the UHS to shed light on the evolution of living standards in cities. In Table E5, we report the estimates from specification (1) with (log) consumer prices as dependent variables. We find that there is an indirect impact of new workers in cities, as consumers of non-tradable goods. These new residents may boost demand for some non-tradable goods, which may benefit firms providing these goods (e.g., housing) or affect firms relying on these goods or services (e.g., with a land-intensive production). We find however that food prices decrease, an effect which is a byproduct of our shift-share design: crop prices affect living costs in neighboring cities both through the arrival of migrants and through the price of

Table E4. Impact of migration inflows on urban residents.

	Wage	Employee	Unemployed	Self-employed
	(1)	(2)	(3)	(4)
Panel A: All urban residents				
Migration	-0.298	-0.041	0.020	0.021
	(0.300)	(0.114)	(0.088)	(0.060)
Observations	187	187	187	187
F-stat. (first stage)	15.29	15.29	15.29	15.29
Panel B: Low-skill urban residents				
Migration	-0.979	-0.079	0.057	0.022
	(0.412)	(0.095)	(0.107)	(0.047)
Observations	187	187	187	187
F-stat. (first stage)	20.21	20.21	20.21	20.21

Notes: Robust standard errors are reported between parentheses. In Panel A, the dependent variable is constructed from the difference in outcomes between 2002 and 2006, for *all* urban residents. In Panel B, we restrict the data construction to urban residents with educational attainment lower than a high school degree. *Wage* is the (log) hourly wage in real terms. *Employee* is a dummy for receiving a wage, while *Self-employed* is a dummy equal to 1 for individuals who are self-employed or employers.

Table E5. Impact of migration inflows on prices in cities.

	CPI	CPI (food)	CPI (non-food)
	(1)	(2)	(3)
Migration	0.327	-0.237	0.564
	(0.141)	(0.102)	(0.216)
Observations	187	187	187
F-stat. (first stage)	15.29	15.29	15.29

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The *CPI*s are the (log) Consumer Price Indices, as computed from the average prices and quantities reported in the UHS survey.

food.

E.4 Complements on production restructuring

The text analysis associates HS 6-digit product codes for (up to) three main products at the establishment level. In Section III, we rely on this classification to observe

changes in the main product and characterize the direction of this change in terms of input. This section provides complements to this analysis and looks at: changes in the human- and physical-capital intensity of products measured through product similarity and the input-output matrix, changes in the technological content of products; changes in the number of products reported by establishments; changes in product similarity for establishments reporting several products.

We use specification (5), in which the causal impact of migration is identified from a long difference in outcomes between the beginning and the end of the period. As in Section III, we clean the estimates from fixed effects at the level of the HS 6-digit code for the main product in 2000. In Table E6, we adopt two alternative ways of classifying products as skill- and capital-intensive. We first use the language proximity score provided by the Natural Language Processing algorithm, and compute for the main product of each firm the average skill- and capital-intensity of all the other firms weighted by the proximity of the HS-6 code of their final product. The results in Panel A of Table E6 are very similar to those of Table 7: following a migration shock firms adopt products that are more similar to those of firms with low-skill and low-capital intensity. Next, we use the input-output matrix, and compute for the main product of each firm the average skill- and capital-intensity of all the other firms weighted by the input-output matrix coefficients. The results in Panel B of Table E6 suggest that firms adopt products that use inputs coming from low-skill low-capital firms, or that are used themselves as inputs by firms with low-skill low-capital production methods.

In Panel A of Table E7, we identify the causal impact of migration on the technological content of products. We compute three measures of technology, all based on the index of technology closeness between different industries developed in Bloom et al. (2013) and based on U.S. patent citations, which we transform into technology closeness between different products, $\{\tau_{p,q}\}_{p,q}$. In the first column, we report changes in the sum of the technology closeness index across industries, $\sum_q \tau_{p,q}$, for an establishment whose main product is p. This variable captures the intensity of technological spillovers for a given product. In the second column, we report changes in the number of links to other industries, i.e., $\sum_{q \neq p} \mathbb{1}_{\tau_{p,q}>0}$. In the third column, we report changes in the Herfindahl index based on technology closeness, i.e., $\sum_{q \neq p} \tau_{p,q}^2 / (\sum_{q \neq p} \tau_{p,q})^2$. These latter two variables capture the width of technology spillovers across industries: a product with a small number of links to other industries or a large Herfindahl index would indicate a "niche technology." Panel A of Table E7 shows that migration pushes establishments towards products with fewer and more concentrated citations across industries. This observation may explain

Table E6. Impact of migration inflows on urban firms—production restructuring—alternative classifications.

Change in product	High ed. (1)	Low ed. (2)	High K/L (3)	Low K/L (4)
Panel A: Using language proximity to characterize products				
Migration	0.025	0.068	0.027	0.067
	(0.034)	(0.026)	(0.029)	(0.023)
Observations	19,189	19,189	19,189	19,189
Panel B: Using input-output links to characterize products				
Migration	-0.006	0.147	0.073	0.069
	(0.023)	(0.036)	(0.029)	(0.022)
Observations	22,151	22,151	22,151	22,151

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. *Migration* is the immigration rate, i.e., the migration flow divided by destination population at baseline. The dependent variables are dummy variables equal to one if there is any change in the main product (1), and if this change goes toward products manufactured by establishments with a more (2) or less (3) educated workforce, and by more (4) or less (5) capital-abundant establishments. See Section I and Equation (5) for a description of the IV specification.

part of the reduction in patenting observed in Section III: following an expansion of labor supply, firms reorient their production lines towards products that are "less reliant on" technological innovation.

In Panel B of Table E7, we identify the causal impact of migration on the number of different products produced by the establishment. Two products are considered to be different if they have different textual descriptions in column 1, different HS 6-digit product codes in column 2, and different HS 3-digit product codes in column 3. We find that migration increases the number of products, even though the effects are quite small and not statistically significant (e.g., a 10% increase in the migration rate increases the number of different HS 3-digit product codes produced by the firm by 0.0077). In Panel C, the sample consists of establishments reporting at least two product codes in 2000 *and* in 2006, and we look at the causal impact of migration on indices of similarity between these products. These indices of similarity are constructed as an average of proximity measures between the different unique pairs of products. We construct these similarity indices from language proximity (see Appendix A.2), the input-output matrix (using input-output accounts in the United States, in 2002, Stewart et al., 2007), and the technology closeness measure

(Bloom et al., 2013). We do not find strong evidence that the similarity changes markedly along any of these three dimensions.

Table E7. Impact of migration inflows on urban firms—technological content, number of products, and similarity between products.

Technological content	Citations	Cross-citations	Herfindahl
	(1)	(2)	(3)
Panel A: Technological content			
Migration	-0.008	-7.723	0.023
	(0.004)	(2.193)	(0.008)
Observations	27,062	27,062	3,681
F-stat. (first stage)	23.05	23.05	20.83
Number of products	Text	HS6	HS3
	(1)	(2)	(3)
Panel B: Number of products			
Migration	0.146	0.077	0.062
	(0.070)	(0.063)	(0.059)
Observations	27,062	25,273	25,273
F-stat. (first stage)	23.05	22.46	22.46
Similarity index	Language	I/O	Technology
	(1)	(2)	(3)
Panel C: Similarity between products			
Migration	0.009	-0.002	-0.036
	(0.019)	(0.003)	(0.022)
Observations	2,815	2,761	2,776
F-stat. (first stage)	23.31	23.04	23.11

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. In Panel A, we rely on the measure of technology closeness between different industries developed in Bloom et al. (2013) and based on U.S. patent citations. The dependent variables are: (1) the sum of technology closeness measures across industries; (2) the sum of technology links with other industries (technology closeness measure different than zero); (3) a Herfindahl index of technology closeness measures across industries. In Panel B, the dependent variables are the number of different products produced by the establishment. Two products are considered to be different if they have different descriptions (column 1), different HS 6-digit product codes (column 2), different HS 3-digit product codes (column 3). In Panel C, the sample is constituted of establishments having at least two products in 2000 and in 2006. The dependent variables are similarity indices based on language proximity (see Appendix A.2), an input-output matrix (using input-output accounts in the United States, in 2002, Stewart et al., 2007), and the technology closeness measure (Bloom et al., 2013).

F Sensitivity analysis

In this Appendix, we investigate the robustness of our results to variations along the different steps of the empirical method. We first assess the sensitivity of the emigration effect to alternative specifications of the shift-share design; we then consider the robustness of our findings to other aspects of the baseline specification (definition of migrant flows, sample selection and outliers, spatial correlation).

F.1 The shift-share design

This section assesses the robustness of our baseline second-stage estimates (specification 5) to the exact construction of the shift-share instrument.

The shift We proceed as in Appendix C.3—which describes the sensitivity of our first stage to the shift-share design—and we consider the following alternative specifications for the shift-share instrument: (i) we use prices from the World Bank Commodities Price Data; (ii) we restrict the set of crops; (iii) we use an AR(2) specification instead of an AR(1); (iv) we isolate innovations in commodity prices with a HP filter. We replicate the results of Table 3 in Table F1 under these different specifications. As apparent from Table F1, our main findings are robust to alternative specifications for the "shift."

The share We show in Panels A and B of Table F2 that our main findings are robust to alternative specifications for the "share." More specifically, we construct the shift-share instrument using (a) previous migrant cohorts, arrived at destination before 1995, and (b) a reallocation of migrants across prefectures along a gravity model, as in Appendix C.3.

F.2 The empirical specification

Definition of migrant flows We show in Table F3 that our main findings are robust to alternative sample choices to define migration spells: all migrant spells, irrespective of their motive (Panel A), males only (Panel B), individuals with secondary education or less (Panel C), migrant spells between prefectures distant of at least 100 km (Panel D), migrant spells between prefectures distant of at least 300 km (Panel E), migration flows not corrected for return migration (Panel F).

Censorship and outliers In the baseline specification (5), we apply a 99% winsorization to firm outcomes and immigration rates. We show in Panels A, B and C

Table F1. Impact of migration inflows on urban firms—sensitivity to the construction of the price shock.

	Labor cost (1)	Employment (2)	K/L ratio (3)	Y/L ratio (4)
Panel A: World Bank "Pink Sheet" prices				
Migration	-0.107	0.305	-0.462	-0.398
	(0.067)	(0.065)	(0.113)	(0.115)
Observations	31,886	31,886	31,886	31,886
Panel B: Restricted list of crops				
Migration	-0.245	0.259	-0.386	-0.515
	(0.076)	(0.047)	(0.088)	(0.116)
Observations	31,886	31,886	31,886	31,886
Panel C: AR(2) specification				
Migration	-0.144	0.298	-0.430	-0.433
	(0.062)	(0.054)	(0.095)	(0.108)
vations	31,886	31,886	31,886	31,886
Panel D: HP-filtered prices				
Migration	-0.199	0.279	-0.436	-0.484
	(0.067)	(0.047)	(0.091)	(0.113)
Observations	31,886	31,886	31,886	31,886

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. In Panel A, the price shock is constructed using World Bank "Pink Sheet" prices; in Panel B, the price shock is constructed using a restricted list of crops (i.e., crops which do not require any inference to match harvested commodities and traded commodities); in Panel C, the price shock is constructed from isolating an innovation with an AR(2) specification (instead of an AR(1) in the baseline); in Panel D, the price shock is constructed using a Hodrick-Prescott filtering (with a parameter of 14,400) on the (log) price of each commodity. See Section I and Equation (5) for a description of the IV specification.

of Table F4 that our main findings are robust to using the actual, uncensored immigration rate, the actual, uncensored firm outcomes, or both.

Spatial correlation The clustering of standard errors in the baseline specification (5) imperfectly deals with the heteroskedasticity induced by the multiplicative structure of shift-share designs and the spatial auto-correlation in shifts (and shares). In this extension, we provide a sensitivity analysis of our main findings (Table 3) relying on the inference results developed in Adão et al. (2019), and the transformation suggested by Borusyak et al. (2018).

Table F2. Impact of migration inflows on urban firms—sensitivity to the construction of the matrix of migration patterns.

	Labor cost	Employment	K/L ratio	Y/L ratio
	(1)	(2)	(3)	(4)
Panel A: Pre-1995 migrant shares				
Migration	-0.182	0.271	-0.410	-0.453
	(0.073)	(0.056)	(0.096)	(0.119)
Observations	31,886	31,886	31,886	31,886
Panel B: Shares predicted from gravity model				
Migration	-0.138	0.296	-0.389	-0.446
	(0.062)	(0.054)	(0.083)	(0.104)
Observations	31,886	31,886	31,886	31,886

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. In Panel A, the matrix of migration patterns is constructed using the stock of migrants at destination in 2000, having arrived in 1995 or before; in Panel B, the matrix of migration patterns is constructed from a gravity model. See Section I and Equation (5) for a description of the IV specification.

We first consider the baseline specification at the level of destinations in Panel A of Table F5, and replace our standard errors clustered at the level of prefectures by: (i) standard errors clustered at the level of the 60 Chinese provinces, (ii) a continuous modeling of the heteroskedasticity across prefectures of destination (following Conley, 1999). We then consider the estimates suggested by Adão et al. (2019): we report the baseline AKM and AKM0 estimates for the standard errors, as well as a specification allowing for clustering.

In Panel B, we apply standard procedures to deal with heteroskedasticity in a transformed specification at origin (Borusyak et al., 2018). We report (i) robust standard errors, (ii) standard errors clustered at the level of the 60 Chinese provinces, (iii) a continuous modeling of the heteroskedasticity across prefectures of origin (following Conley, 1999).

Table F3. Impact of migration inflows on urban firms—sensitivity to the definition of migrant spells and migration patterns.

	Labor cost (1)	Employment (2)	K/L ratio (3)	Y/L ratio (4)
Panel A: all migrants				
Migration	-0.147	0.294	-0.431	-0.437
	(0.062)	(0.053)	(0.095)	(0.108)
Observations	31,886	31,886	31,886	31,886
Panel B: male migrants				
Migration	-0.273	0.565	-0.846	-0.825
	(0.121)	(0.104)	(0.192)	(0.211)
Observations	31,886	31,886	31,886	31,886
Panel C: low-education migrants				
Migration	-0.170	0.342	-0.504	-0.513
	(0.072)	(0.067)	(0.120)	(0.135)
Observations	31,886	31,886	31,886	31,886
Panel D: Migrant spells $>$ 100 km				
Migration	-0.136	0.281	-0.437	-0.421
	(0.061)	(0.049)	(0.100)	(0.107)
Observations	31,886	31,886	31,886	31,886
Panel E: Migrant spells $>$ 300 km				
Migration	-0.001	0.263	-0.602	-0.395
	(0.107)	(0.066)	(0.185)	(0.157)
Observations	31,886	31,886	31,886	31,886
Panel F: No adjustment for return migration				
Migration	-0.168	0.336	-0.493	-0.500
	(0.070)	(0.063)	(0.112)	(0.126)
Observations	31,886	31,886	31,886	31,886

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. In Panel A, all migration spells are considered; we restrict the sample to males only in Panel B, and to low-education individuals in Panel C; we consider migration spells between prefectures distant of at least 100 km in Panel D (resp. 300 km in Panel E); we do not adjust the migration flows for return migration in Panel F. See Section I and Equation (5) for a description of the IV specification.

Table F4. Impact of migration inflows on urban firms—sensitivity to outliers.

	Labor cost (1)	Employment (2)	K/L ratio (3)	Y/L ratio (4)
Panel A: No winsorizing of firm outcomes				
Migration	-0.146	0.341	-0.422	-0.460
	(0.064)	(0.059)	(0.098)	(0.109)
Observations	31,886	31,886	31,886	31,886
Panel B: No winsorizing of migration rates				
Migration	-0.079	0.158	-0.232	-0.235
	(0.042)	(0.057)	(0.088)	(0.096)
Observations	31,886	31,886	31,886	31,886
Panel C: No winsorizing (outcomes or migration rates)				
Migration	-0.078	0.183	-0.227	-0.247
	(0.043)	(0.062)	(0.090)	(0.093)
Observations	31,886	31,886	31,886	31,886

Notes: Standard errors are clustered at the prefecture level and reported between parentheses. The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. In Panel A, firm outcomes are not winsorized, but migrant flows are. In Panel B, migrant flows are not winsorized, but firm outcomes are. In Panel C, neither firm outcomes nor migrant flows are winsorized. See Section I and Equation (5) for a description of the IV specification.

Table F5. Impact of migration inflows on urban firms—alternative standard errors.

	Labor cost (1)	Employment (2)	K/L ratio (3)	Y/L ratio (4)
Panel A: Destination (firm-level)				
Coefficient on migration	-0.147	0.295	-0.432	-0.438
S.E. Prefecture Cluster	0.062	0.053	0.095	0.109
S.E. Province Cluster	0.068	0.040	0.078	0.071
S.E. Conley (300km radius)	0.059	0.044	0.072	0.047
S.E. AKM (1)	0.030	0.020	0.025	0.028
S.E. AKM (0)	0.031	0.021	0.027	0.029
S.E. AKM (Province Cluster)	0.057	0.036	0.056	0.041
Panel B: Origin (prefecture-level)				
Coefficient on Migration	-0.147	0.294	-0.431	-0.437
S.E. Robust (baseline)	0.029	0.019	0.024	0.026
S.E. Province Cluster	0.056	0.035	0.055	0.040
S.E. Conley (300km radius)	0.058	0.030	0.056	0.040

Notes: The sample is composed of the firms present every year in the NBS firm census between 2000 and 2006. See Section I and Equation (5) for a description of the IV specification.

References

Abebe, Girum, Stefano Caria, Marcel Fafchamps, Paolo Falco, Simon Franklin, and Simon Quinn, "Anonymity or Distance? Job Search and Labour Market Exclusion in a Growing African City," CSAE Working Paper Series 2016.

Adão, Rodrigo, Michal Kolesár, and Eduardo Morales, "Shift-share designs: Theory and inference," *The Quarterly Journal of Economics*, 2019, *134* (4), 1949–2010.

Akcigit, Ufuk, Salome Baslandze, and Stefanie Stantcheva, "Taxation and the international mobility of inventors," *American Economic Review*, 2016, *106* (10), 2930–81.

Alfonsi, Livia, Oriana Bandiera, Vittorio Bassi, Robin Burgess, Imran Rasul, Munshi Sulaiman, and Anna Vitali, "Tackling Youth Unemployment: Evidence from a Labour Market Experiment in Uganda," Technical Report 64, STICERD, LSE December 2017.

Autor, David H, David Dorn, and Gordon H Hanson, "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, 2013, *103* (6), 2121–68.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen, "Identifying technology spillovers and product market rivalry," *Econometrica*, 2013, *81* (4), 1347–1393.

Boehm, Johannes and Ezra Oberfield, "Misallocation in the Market for Inputs: Enforcement and the Organization of Production," Technical Report, National Bureau of Economic Research 2018.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel, "Quasi-experimental shift-share research designs," Technical Report, National Bureau of Economic Research 2018.

Boustan, Leah Platt, Price V. Fishback, and Shawn Kantor, "The Effect of Internal Migration on Local Labor Markets: American Cities during the Great Depression," *Journal of Labor Economics*, October 2010, *28* (4), 719–746.

Brandt, Loren, Johannes Van Biesebroeck, and Yifan Zhang, "Challenges of working with the Chinese NBS firm-level data," *China Economic Review*, 2014, *30* (C), 339–352.

Bryan, Gharad and Melanie Morten, "The aggregate productivity effects of internal migration: Evidence from indonesia," *Journal of Political Economy*, 2019, *127* (5), 2229–2268.

Buera, Francisco J, Joseph P Kaboski, and Yongseok Shin, "Finance and development: A tale of two sectors," *The American Economic Review*, 2011, *101* (5), 1964–2002.

Card, David, "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration," *Journal of Labor Economics*, 2001, *19* (1), 22–64.

Chen, Shuai, Paulina Oliva, and Peng Zhang, "The effect of air pollution on migration: evidence from China," Technical Report, National Bureau of Economic Research 2017.

Conley, Timothy G, "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

Davis, Steven J and John C Haltiwanger, "Job creation and destruction," *MIT Press Books*, 1998, *1.*

Demurger, Sylvie, Marc Gurgand, Shi Li, and Ximing Yue, "Migrants as second-class workers in urban China? A decomposition analysis," *Journal of Comparative Economics*, December 2009, *37* (4), 610–628.

Dustmann, Christian and Albrecht Glitz, "How Do Industries and Firms Respond to Changes in Local Labor Supply?," *Journal of Labor Economics*, 2015, *33* (3), 711 – 750.

Ebenstein, Avraham and Yaohui Zhao, "Tracking rural-to-urban migration in China: Lessons from the 2005 inter-census population survey," *Population Studies*, 2015, *69* (3), 337–353.

Fan, Cindy C., *China on the Move*, Routledge, 2008.

Feng, Shuaizhang, Yingyao Hu, and Robert Moffitt, "Long run trends in unemployment and labor force participation in urban China," *Journal of Comparative Economics*, 2017, *45* (2), 304 – 324.

Freeman, Richard, Wenquan Liang, Ran Song, and Christopher Timmins, "Willingness to pay for clean air in China," *Journal of Environmental Economics and Management*, 2019, *94*, 188–216.

Ge, Suqin and Dennis Tao Yang, "Changes In China's Wage Structure," *Journal of the European Economic Association*, 04 2014, *12* (2), 300–336.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift, "Bartik Instruments: What, When, Why, and How," *American Economic Review*, forthcoming.

Hsieh, Chang-Tai and Peter J. Klenow, "Misallocation and Manufacturing TFP in China and India," *The Quarterly Journal of Economics*, November 2009, *124* (4), 1403–1448.

Jaeger, David A, Joakim Ruist, and Jan Stuhler, "Shift-Share Instruments and the Impact of Immigration," CEPR Discussion Papers 12701, C.E.P.R. Discussion Papers February 2018.

Lewis, Ethan, "Immigration, Skill Mix, and Capital Skill Complementarity," *The Quarterly Journal of Economics*, 2011, *126* (2), 1029–1069.

Li, Shen, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du, "Analogical Reasoning on Chinese Morphological and Semantic Relations," in "Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)" Association for Computational Linguistics 2018, pp. 138–143.

Midrigan, Virgiliu and Daniel Yi Xu, "Finance and misallocation: Evidence from plant-level data," *The American Economic Review*, 2014, *104* (2), 422–458.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in "Advances in neural information processing systems" 2013, pp. 3111–3119.

Monras, Joan, "Immigration and Wage Dynamics: Evidence from the Mexican Peso Crisis," Working Papers hal-01127022, HAL March 2015.

Monte, Ferdinando, Stephen J Redding, and Esteban Rossi-Hansberg, "Commuting, migration, and local employment elasticities," *American Economic Review*, 2018, *108* (12), 3855–90.

Munshi, Kaivan, "Networks in the modern economy: Mexican migrants in the US labor market," *The Quarterly Journal of Economics*, 2003, *118* (2), 549–599.

Oberfield, Ezra and Devesh Raval, "Micro data and macro technology," 2014.

Park, Albert, "Rural-urban inequality in China," in Shahid Yusuf and Karen Nabeshima, eds., *China Urbanizes: Consequences, Strategies, and Policies*, The World Bank, 2008.

Song, Yang, "What should economists know about the current Chinese hukou system?," *China Economic Review*, 2014, *29*, 200–212.

Stewart, Ricky L, Jessica Brede Stone, and Mary L Streitwieser, "US benchmark input-output accounts, 2002," *Survey of Current Business*, 2007, *87* (10), 19–48.

Tombe, Trevor and Xiaodong Zhu, "Trade, Migration, and Productivity: A Quantitative Analysis of China," *American Economic Review*, May 2019, *109* (5), 1843–72.

Zhang, Li and Li Tao, "Barriers to the acquisition of urban hukou in Chinese cities," *Environment and Planning A*, 2012, *44* (12), 2883–2900.