Version of Record: https://www.sciencedirect.com/science/article/pii/S0304387819301142
Manuscript_be4bddfff188256f29439abc7993d148

Unequal Migration and Urbanisation Gains in China [1]

Pierre-Philippe Combes [2] Sylvie Démurger [3] Shi Li [4] Jianguo Wang [5]

January 17, 2019

Abstract

We assess the role of internal migration and urbanisation in China on the nominal earnings of three groups of workers (rural migrants, low-skilled natives, and high-skilled natives). We estimate the impact of many city and city-industry characteristics that shape agglomeration economies, as well as migrant and human capital externalities and substitution effects. We also account for spatial sorting and reverse causality. Location matters for individual earnings, but urban gains are unequally distributed. High-skilled natives enjoy large gains from agglomeration and migrants at the city level. Both conclusions also hold, to a lesser extent, for low-skilled natives, who are only marginally negatively affected by migrants within their industry. By contrast, rural migrants slightly lose from migrants within their industry while otherwise gaining from migration and agglomeration, although less than natives. The different returns from migration and urbanisation are responsible for a large share of wage disparities in China.

JEL Codes: O18, R12, R23, J31, O53.

Keywords: urban development; agglomeration economies; wage disparities; migrants; human capital externalities; China.

[1] We are grateful to Laurent Gobillon for in-depth discussions about the simultaneous identification of agglomeration and human capital externalities and to Vernon Henderson for suggesting further instruments for migrant variables. We thank the many seminar participants who offered helpful comments at various stages of the paper. We also thank two anonymous referees and the Editor of the journal. Any remaining mistakes are our own. Financial support from the Aix-Marseille School of Economics, the French Centre National de la Recherche Scientifique (CNRS) International Associated Laboratory LIA CHINEQ, the Agence Nationale de la Recherche (ANR) research programs ANR-14-ORAR-0002-01 and ANR-18-CE41-0003-02, the National Natural Science Foundation of China (Grant 71503023) and the Beijing Social Science Foundation (Grant 16YJC052) is gratefully acknowledged.

[2] Univ Lyon, CNRS, GATE UMR 5824, F-69130 Ecully, France; Sciences Po, Department of Economics, 28, Rue des Saints-Pères, 75007 Paris, France. Research fellow at CEPR. Email: ppcombes@gmail.com.

[3] Corresponding author. Univ Lyon, CNRS, GATE UMR 5824, F-69130 Ecully, France. Research fellow at IZA. Email: sylvie.demurger@cnrs.fr.

[4] Business School, Beijing Normal University, China. Research fellow at IZA. Email: lishi@bnu.edu.cn.

[5] Beijing Information Science and Technology University, China. Email: jgwang0225@gmail.com.

1 Introduction

Rising earnings inequality is a recurring concern in the policy debate in China. A sizeable body of literature highlights individual and firm characteristics as being important determinants of the unequal distribution of workers' income in China[1]. Simultaneously, 56.1% of the Chinese population lives in cities as of 2015, having risen from only 26.4% in 1990, with a migrant population share that amounts to 18% (close to 250 million people), against only 1.9% in 1990. Wage dispersion across cities is also large. From the 2005 China Population Survey, one can compute that the average wage in 2005 in the city at the ninth decile is 70% higher than the average wage in the city at the first decile. Despite the accelerating urbanisation and large spatial wage disparities, the role that migration and urbanisation play in determining the earnings of both urban natives and rural migrants in Chinese cities has barely been studied. This is the purpose of the present paper.

Different strands of literature investigate how city externalities impact workers' earnings beyond the role of workers' own individual characteristics. A large literature has investigated the role of the human capital effects created by the presence of more skilled workers in cities (see Moretti, 2004, for a review) and emphasises the difficulty of identifying human capital externalities separately from substitution effects within the production function for workers of different types. This literature often assigns a minor role to agglomeration effects, either at the overall city level or at the industry level within the city (the city-industry level). Such agglomeration effects are, conversely, central in urban economics (see Combes and Gobillon, 2015, for a review) that quantifies the impact on wages of city size and a number of variables that relate to access to distant markets, or, at the city-industry level, the role of specialisation – defined as the share of the worker's industry in city employment. Similarly, the literature on the role of migrants in cities focuses primarily on international migration and the substitution effects between natives and migrants – especially in the case of Mexican workers in the US – but rarely simultaneously assesses the role of other city characteristics (see Lewis and Peri, 2015, for a review).

The aim of this paper is to contribute to all three literatures by proposing a general framework that encompasses the various sources of local externalities and substitution effects to evaluate the unequal wage gains across workers that arise from them. We assess the role of both city and city-industry characteristics that capture human capital, agglomeration, and migrant impacts on individual earnings in a single analytical framework, from which we derive and estimate an empirical specification. We also explore whether the gains are similar across workers, or whether location and industry choices can themselves be a source of disparities among heterogeneous workers. To the

[1]Overviews are provided in Li and Sicular (2014); Wang, Wan and Yang (2014); Xie and Zhou (2014).

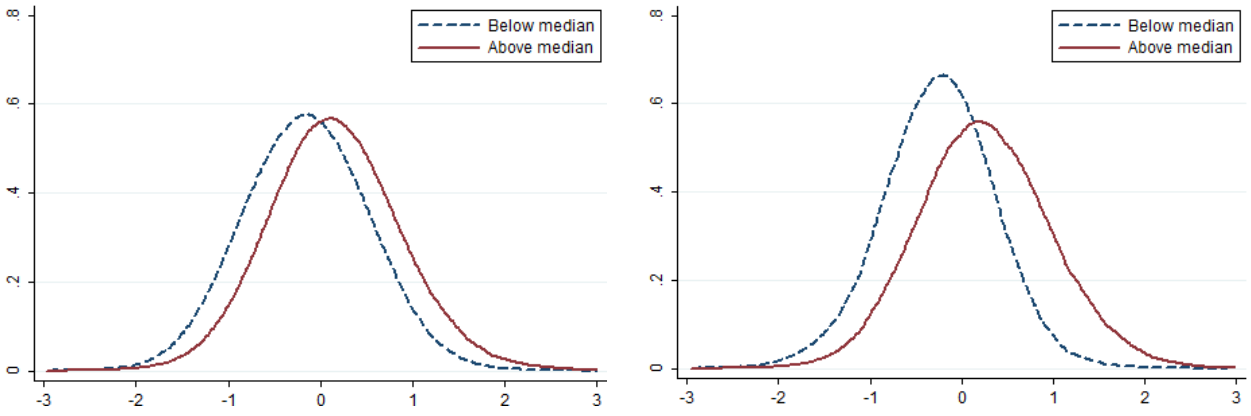
extent that different types of workers correspond to different inputs in the production function, there is no reason why the impact of city and city-industry characteristics that shape local externality and substitution effects should be the same across workers. This might typically be the case in China, where workers differ not only along a skill dimension but also according to their migration status. Moreover, even if externality and substitution effects cannot be identified separately, we expect substitution effects to be stronger relative to externalities at the city-industry level. We test for this by assessing the role of city and city-industry characteristics simultaneously and comparing their magnitude. Understanding whether and how the different groups of workers benefit from spatial concentration and from the presence of migrants and high-skilled workers in cities has key policy implications for China's future urban growth, including regarding whether to accelerate or to slow down its tremendous urbanisation process.

The framework we rely on derives inverse labour demand from a local production function, which relates nominal wages to individual, city-industry, and city characteristics. We focus on nominal wages and do not intend to assess whether these characteristics affect individual welfare generally. Studying real income disparities would require evaluating the cost of living in the city and the value of city amenities, and both should be worker's type specific. A few papers have begun to provide such an assessment for the US (see for instance Albouy, 2008; Moretti, 2013), but the task remains, for the present, infeasible for China because of data limitations. In our framework, local prices for the cost of living do not intervene, and as the urban economics literature and our model below emphasise, the price of the goods sold does so only as a part of the gains from (or costs of) spatial concentration. Quantifying the city-level determinants of nominal earnings remains important *per se* and is also a preliminary and necessary step before studying real income disparities.

Figure 1 illustrates one of our main results, the relationship between the migrant population in cities and the wage structure. Using raw data, Figure 1 plots the distribution of wages in cities below and above the median for the share of migrant workers present in the city, separately for low-skilled urban natives (Panel (a)) and for high-skilled urban natives (Panel (b)). Both panels show that the distribution of wages in cities above the migrant share median is right-shifted compared to the distribution in cities below the median: Urban native workers have higher wages when they locate in cities with more migrants. The comparison between Panel (a) and Panel (b) further suggests a larger effect for high-skilled workers compared to low-skilled workers.

However, making a causal interpretation of such a migrant impact – as well as that of any other determinant of city externalities – requires us to address two major issues: *i*) the non-random spatial sorting of workers, for example, if more-able workers with higher wages are over-represented

Figure 1 – Distribution of the (logarithm of) individual wages for low- and high-skilled urban natives in cities above and below the median share of migrants



Panel (a): Low-skilled workers

Panel (b): High-skilled workers

Notes: Data drawn from the 2005 China Population Survey. See Section 2.3 and Appendix B.

in some cities (typically larger cities and those with more migrants, for instance.) and *ii*) reverse causality from wages to city characteristics if the location choices of firms and workers are driven by spatial wage differences. We address the first concern by using individual data to control for individual characteristics. To treat endogeneity, we propose a wider set of instruments than the existing literature on China, using past population censuses dating back to 1964 and past migration flows predicted by a gravity model.

We find that location matters for Chinese workers' nominal earnings and that urban gains are unequally distributed across workers. We are also able to identify the impact of migration simultaneously at the city-industry and the city levels, and we show that it differs between these levels. The presence of migrants at the city level has the largest impact for all groups of workers, including migrants. Increasing the migrant ratio to low-skilled workers benefits high-skilled urban natives the most, followed by low-skilled urban natives, and finally, rural migrants, albeit to a lesser extent. Our estimates suggest for instance that everything else constant, the higher migrant ratio in Shanghai leads to high-skilled urban natives' wages being 44% higher there than in Nanchang[2]. The corresponding Shanghai-Nanchang gap due to the presence of migrants reduces to 26% for low-skilled urban natives and to only 10% for rural migrants. In contrast to the large city-level impact of migration, we further show that, at the city-industry level, both low-skilled urban natives and rural migrants tend to be negatively affected by migrants within their industry, indicating that the substitution effect may counterbalance any positive external effect of migration within an

[2]Nanchang is the capital city of Jiangxi province, located 700 kilometers west of Shanghai. The employment density in both cities is of a similar magnitude, but the cities differ greatly in terms of their migrant ratio to low-skilled workers, which is more than 6 times larger in Shanghai.

industry. However, the marginal negative impact is much smaller in magnitude than the positive one at the city level. All workers are also found to benefit from standard agglomeration effects, notably city size, and we again observe a decreasing magnitude from high-skilled to low-skilled and then to migrant workers. For instance, the wage gain of moving from the first quartile to the last quartile of the city employment density is more than twice smaller, at 3%, for rural migrants than for natives (at around 7%). We document a number of other city and city-industry gains that also differ across workers. Finally, a Oaxaca-Blinder decomposition exercise suggests that the different returns to city characteristics and the fact that the three groups of workers work in different cities are responsible for a large share of average wage gaps in China.

As summarised by Chauvin, Glaeser, Ma and Tobio (2017), most of the empirical literature on agglomeration effects focuses on developed countries (especially on the US), and little is known about the impact of urbanisation in developing countries despite the global importance of the phenomenon in these countries. Chauvin et al. (2017) also argue that only some of the stylised facts documented about cities in the US apply to cities of the developing world, and they call for more research on cities in the developing world. Moreover, differences in the levels of technology, mobility, and trade costs, even if they are due only to the larger share of manufacturing with respect to the services industry in developing world cities, may all imply varying magnitudes for city externality and substitution effects. A burgeoning literature examines city effects in China, specifically regarding the impact of migrants (Meng and Zhang, 2010; de Sousa and Poncet, 2011; Combes, Démurger and Li, 2015; Han and Li, 2017), the gains from larger cities (Au and Henderson, 2006; Combes et al., 2015) or market access (Hering and Poncet, 2010), and human capital externalities (Liu, 2014). However, none of these papers jointly considers all local externalities at both the city and city-industry levels or allows for an heterogeneous impact across workers' types. The geography considered encompasses either fewer cities or a much larger scale (the province rather than city level), and in general, these works do not seek to propose a comprehensive treatment of both spatial sorting and endogeneity as we do here. Combes et al. (2015) provide preliminary answers to the role of city effects in China and show that location, and migrants in particular, matter for urban resident workers. However, given the limited coverage of the data set they use, they focus primarily on a few overall city characteristics, in 83 cities only, and they do not study how wage gains differ between local residents and migrant workers. We overcome these limitations by using the 2005 China Population Survey. This large dataset allows us to investigate simultaneously the role of many more local characteristics, computed at both the city and city-industry levels and instrumented by a larger set of variables, on individual earnings in 257 cities.

The remainder of the paper is organised as follows. Section 2 describes the theoretical mechanisms involved and our empirical strategy. Section 3 provides a variance analysis that highlights the respective roles of individual, city-industry, and city characteristics on individual wages and then proceeds with our estimations on unequal urban gains. Section 4 concludes the paper.

2 Mechanisms, empirical strategy and data

2.1 Production function and urban externalities

Our objective is to evaluate the impact of migration and urbanisation on individual nominal wages for three groups of workers that we treat as imperfect substitutes in the production function. This function also encompasses city and city-industry externalities from agglomeration, migrants, and human capital. We derive a specification that corresponds to an (inverse) labour demand function. It highlights the role of migrants, human capital, and a set of other urban characteristics, such as overall size and market access. These effects are usually assessed separately in their respective literature, with no consideration of any possible inter-dependencies (see Combes and Gobillon, 2015, for a review). The model we sketch here allows us to disentangle the local mechanisms at work at both the city and city-industry levels, and to discuss endogeneity issues and the sources of identification we use.

Our empirical strategy is based on a production function in which output in city c and industry s, Y_{cs}, is produced using a labour input composed of high-skilled workers, H_{cs}, low-skilled workers, L_{cs}, and migrants, M_{cs}. Migrants are assumed to be substitute for high- and low-skilled workers but to different extents. Typically, the main mechanisms can be grasped by considering a CES production function:

$$Y_{cs}\left(H_{cs}, L_{cs}, M_{cs}\right) = \left[\left(A_{cs}^{H} H_{cs}\right)^{\rho} + \left[\left(A_{cs}^{L} L_{cs}\right)^{\eta} + \left(A_{cs}^{M} M_{cs}\right)^{\eta}\right]^{\frac{\rho}{\eta}}\right]^{\frac{1}{\rho}}, \tag{1}$$

where A_{cs}^{k} is the efficiency of workers with skills k ($k = H, L, M$) operating in city c and industry s, and ρ and η are parameters between 0 and 1. As migrants should be more substitute to low-skilled workers than to high-skilled ones, we also expect η to be larger than ρ, so overall $0 < \rho \leq \eta \leq 1$. With individual data that allow us to control for individual skills, we can explicitly measure labour in terms of efficient units while assuming perfect substitution within groups, such that:

$$H_{cs} = \sum_{\substack{\text{high-skilled } i \\ \in \{cs\}}} s_i \ell_i \,, \qquad L_{cs} = \sum_{\substack{\text{low-skilled } i \\ \in \{cs\}}} s_i \ell_i \,, \qquad M_{cs} = \sum_{\substack{\text{migrants } i \\ \in \{cs\}}} s_i \ell_i \,, \tag{2}$$

with ℓ_i being individual i's number of hours worked and s_i the number of efficient labour units per hour. It is possible to solve for wages at the individual level using the first-order conditions that determine the optimal use of each type of labour under perfect competition. For instance, the nominal wage of high-skilled workers, w_i^H, is obtained as:

$$w_i^H = s_i \left(p_{cs} A_{cs}^H\right)^{\rho} \left[\left(p_{cs} A_{cs}^H S_{cs}^H\right)^{\rho} + \left[\left(p_{cs} A_{cs}^L\right)^{\eta} + \left(p_{cs} A_{cs}^M S_{cs}^M\right)^{\eta}\right]^{\rho/\eta}\right]^{\frac{1-\rho}{\rho}} \left(S_{cs}^H\right)^{-(1-\rho)} \equiv s_i \, w_{cs}^H \,, \quad (3)$$

where $S_{cs}^H = \frac{H_{cs}}{L_{cs}}$ and $S_{cs}^M = \frac{M_{cs}}{L_{cs}}$ are the ratios of the number of high-skilled and migrant workers respectively to the number of migrants in industry s, and p_{cs} is the firms' income per unit of good produced in the city-industry (e.g., its price net of intermediate consumption and transport costs to final markets). Similar expressions for low-skilled and migrant workers' wages are given in Appendix A.

We assume that the city-industry component of labour efficiency for each group of workers is a function of both city and city-industry externalities specific to the worker's group k (*City Externalities*$_c^k$ and *City-Industry Externalities*$_{cs}^k$, respectively):

$$p_{cs} A_{cs}^k = \textit{City Externalities}_c^k \times \textit{City-Industry Externalities}_{cs}^k \quad \text{for} \quad k = H, L, M. \quad (4)$$

These externalities are then assumed to be shaped by a set of city and city-industry characteristics. Let us first assume that they are affected only by the main three variables studied in the literature: the size of the city -measured by city total employment, $E_c = H_c + L_c + M_c$-, the share of high-skilled workers in employment, and the share of migrants in employment. Moreover, the latter two variables effects are assumed to arise both within the city and within industries in the city, i.e., to depend on both S_{cs}^H and $S_c^H = \frac{H_c}{M_c}$ and S_{cs}^M and $S_c^M = \frac{M_c}{L_c}$.[3] These assumptions typically summarise into:

$$p_{cs} A_{cs}^k = E_c^{\beta^k} \left(S_c^H\right)^{\mu^k} \left(S_c^M\right)^{\varphi^k} \left(S_{cs}^H\right)^{\lambda^k} \left(S_{cs}^M\right)^{\psi^k} \nu_{cs}^k \quad \text{for} \quad k = H, L, M, \quad (5)$$

where β^k, μ^k, φ^k, λ^k, and ψ^k are group-k-specific parameters that are positive when the corresponding city or city-industry characteristic enhances labour productivity and ν_{cs}^k is a labour efficiency random component at the city-industry level. Combes and Gobillon (2015) detail the

[3] We specify local externatities as a function of the ratio of high-skilled and migrant workers over low-skilled workers, but one could also consider the shares of high-skilled and migrant workers in total employment. This corresponds to a change of variables and a functional form choice that lead to more cumbersome computations and formulas, in particular when the simultaneous role of externalities at both city and city-industry levels is considered as we propose here. However, the intuition is the same.

micro-foundations of these effects. Some mechanisms directly affect labour efficiency, A_{cs}^k , while others operate through the firms' price, p_{cs}. Other mechanisms could go through the cost of inputs other than labour not considered here. Importantly, the parameters correspond to the total net externality effect of each local characteristic. Negative externalities can partly erode positive ones, for instance, or some may affect prices, while others affect labour efficiency. β^k, μ^k, φ^k, λ^k, and ψ^k reflect the resultant of all of the effects at work, which cannot be separately identified. This is a standard feature of the empirical agglomeration literature based on wage equations.

Using both (3) and (5), we can relate the spatial (or time) variations in the local component of wages, w_{cs}^H in equation (3), to the corresponding variations in local characteristics:

$$\begin{aligned}
\frac{dw_{cs}^H}{w_{cs}^H} &= \left[\beta^H - (1-\rho)\left[\left(1-\theta^H-\theta^M\right)\left(\beta^H-\beta^L\right)+\theta^M(\beta^H-\beta^M)\right]\right]\frac{dE_c}{E_c} \\
&+ \left[\mu^H - (1-\rho)\left[\left(1-\theta^H-\theta^M\right)\left(\mu^H-\mu^L\right)+\theta^M(\mu^H-\mu^M)\right]\right]\frac{dS_c^H}{S_c^H} \\
&+ \left[\varphi^H - (1-\rho)\left[\left(1-\theta^H-\theta^M\right)\left(\varphi^H-\varphi^L\right)+\theta^M(\varphi^H-\varphi^M)\right]\right]\frac{ds_c^M}{s_c^M} \qquad (6) \\
&+ \left[\lambda^H - (1-\rho)\left[\left(1-\theta^H-\theta^M\right)\left(\lambda^H-\lambda^L\right)+\theta^M(\lambda^H-\lambda^M)+1-\theta^H\right]\right]\frac{dS_{cs}^H}{S_{cs}^H} \\
&+ \left[\psi^H - (1-\rho)\left[\left(1-\theta^H-\theta^M\right)\left(\psi^H-\psi^L\right)+\theta^M(\psi^H-\psi^M)-\theta^M\right]\right]\frac{ds_{cs}^M}{s_{cs}^M},
\end{aligned}$$

where θ^H (θ^M, respectively) is the share of high-skilled (migrant, respectively) workers in the local wage bill, $\theta^k = \frac{w_{cs}^k\, k_{cs}}{w_{cs}^M\, M_{cs} + w_{cs}^L\, L_{cs} + w_{cs}^H\, H_{cs}}$ for $k = H, M$.

Importantly, equation (6), and the corresponding equations for low-skilled and migrant workers provided in Appendix A, directly match the specifications we estimate (see below), and show that the impact of any city or city-industry characteristic on wages is a function not only of the externality parameter for the corresponding group, but also of the externality parameter of the other two groups, of the elasticities of substitution between groups, and of the share of each group in the local wage bill. Three conclusions are important to keep in mind in terms of identified mechanisms. First, even in the case of positive total net externalities ($\beta^k > 0$, $\mu^k > 0$, $\varphi^k > 0$), if the externalities for a group of workers are larger than for the other group, the impact of a variable on the former group's wage can be negative because of the substitution effect. Indeed, the increase in productivity due to the externality gives incentives to firms to hire relatively more of these workers, which in turn reduces their wage. Second, and by contrast, if groups are perfect substitute ($\rho = \eta = 1$) or if externalities have the same magnitude for all groups ($\beta^H = \beta^L = \beta^M$ or $\mu^H = \mu^L = \mu^M$), wage elasticities for city characteristics are close to the externality effect (β^k, μ^k, φ^k, λ^k, and ψ^k). Finally, regarding

the impact of the share of high-skilled and migrant workers at the city-industry level, S_{cs}^{H} and S_{cs}^{M}, a further negative substitution mechanism arises. For instance, as regards S_{cs}^{H}, the substitution effect reduces the externality gain for high-skilled workers and can even lead to a negative variation in the high-skilled workers' wage following an increase in their share at the city-industry level. Reversely, the substitution effect adds up to the human capital externality effect as regards low-skilled and migrants' wages.

The fact that the externality and substitution effects cannot be identified separately within the impact of the workers' share variable is emphasised in the literature on human capital (see Moretti, 2004). We show that this concern is reinforced when externalities differ in magnitude across workers' types and that it holds for any city or city-industry variable. While keeping this issue in mind is crucial for a correct interpretation of the estimated parameters, it is not the one that we attempt to address here. We only wish to evaluate whether the overall effect of a higher share of a group of workers (high-skilled and migrant workers here) is positive or negative and whether it varies across groups of workers, which in turn would impact the wage gap between groups. Moreover, one can still interpret a positive impact of a variable as indicating that positive externality effects dominate substitution effects when the latter are negative, or that the difference between these two families of effects is larger or lower for a group of workers compared to another. These are important conclusions from a policy perspective, for instance with respect to the city characteristics that would enhance labour productivity and would affect wage gaps between workers.

Interestingly, the ratio of migrants over low-skilled workers is the typical variable considered in the literature on migrants' role in natives' outcomes (see Lewis and Peri, 2015, for a review), which, conversely, typically ignores human capital and other urban externalities. Similar to the ratio of high-skilled workers, the ratio of migrants captures both an *a priori* positive externality effect, arising for instance from the fact that recently arrived migrants take occupations that lead to the lowest returns thus pushing other workers to more rewarding ones, and a substitution effect. This substitution effect is now positive for high-skilled workers and negative for migrants themselves. Regarding low-skilled workers, the substitution effect could be negative if migrants are strong substitutes for them (η much larger than ρ, see Appendix A) or positive in the reverse case. Therefore, we expect the overall impact of migrants to be the largest for high-skilled workers and the smallest for migrants, at least to the extent that the externality effect is similar across groups. When it is not the case, further substitution effects similar to those evoked above occur.

Finally, equation (6) shows that the role of the high-skilled and migrant ratios can be identified at the city and city-industry levels simultaneously. This is one of the advantages of our strategy, and

to the best of our knowledge, this has not previously been considered in the literature. However, the externality and substitution effects interfere at both levels. Nevertheless, we expect substitution effects to operate mostly at the most micro level (the city-industry) and, conversely, to be smaller at the city level. We view high-skilled, low-skilled, and migrant workers as actually substitute inputs at the city-industry level but more as given to the firms at the city-level, as specified in equations (1) and (5).

The urban literature suggests that agglomeration externalities are affected by many more variables than the city's total employment considered above. City total employment is decomposed into the role of the intensive margin, the number of employees per square meter (i.e., density ($Density_c$)) and its extensive margin, the city land area ($Area_c$). Since trade takes place between locations and because workers are mobile, the literature also emphasises the role of access to distant (large) markets that can generate 'imported external economies' in cities (Head and Mayer, 2004). We use two variables to disentangle the role of internal and international markets. The first variable ($Access_c$) is market access à la Harris (1954), which assesses how close the city is to other large Chinese cities. As a large share of Chinese exports moves through coastal ports, the second variable ($Port_c$) measures the city's proximity to the closest seaport, which proxies for access to international markets. We consider the role of a last city variable, reflecting the industrial diversity of the city, $Diversity_c$, which is the inverse of a Herfindhal index based on city-industry shares in city employment. This assesses whether a homogeneous distribution of employment across local industries enhances local productivity, following the initial intuition associated with Jacobs (1969). At the city-industry level, beyond the high-skilled and migrant ratios over low-skilled workers, we also consider the share of workers involved in industry s in city c, reflecting the city's specialisation in this industry ($Specialisation_{cs}$). This is the most studied city-industry variable in urban economics. A higher share of an industry in city employment should favour a number of externalities operating within the industry as suggested in an early contribution by Marshall (1890). Importantly, the fact that the different sources of positive and negative externality and the substitution effects cannot be separately identified applies to any city or city-industry characteristic. The mathematical definition of all variables is provided in Appendix B, Table 4.

2.2 Endogenous location choices and econometric issues

The Mincerian equation augmented by city and city-industry variables given in (7) is the direct empirical equivalent of equation (6) for high-skilled workers and its corresponding equations for

low-skilled and migrant workers reported in Appendix A:

$$\log w_i \;=\; X_i\,\zeta^k + \;\textit{City-Industry Variables}_{cs}\;\delta^k \;+\; \textit{City Variables}_c\;\gamma^k \;+\nu^k_{cs}+\varepsilon^k_i, \tag{7}$$

for $k = H, L, M$, when individual efficiency s_i is specified as:

$$s_i \;=\; X_i\,\zeta^k + \varepsilon^k_i, \tag{8}$$

and *City-Industry Variables*$_{cs}$ and *City Variables*$_c$ are the vectors of city and city-industry characteristics that determine *City Externalities*$_c$ and *City-Industry Externalities*$_{cs}$ in specification (4). As detailed above, seven city characteristics (*Density*$_c$, *Area*$_c$, *Access*$_c$, *Port*$_c$, *Diversity*$_c$, S^H_c, and S^M_c) and three city-industry characteristics (*Specialisation*$_{cs}$, S^H_{cs}, and S^M_{cs}) are considered. ζ^k, δ^k, and γ^k are vectors of group-k parameters that capture the overall externality and substitution effects discussed in Section 2.1. Equation (6) also emphasises that the marginal impact of the variables are *a priori* city-industry specific as they also depend on the labour shares in the local wage bill, θ^k. Standardly, the parameters estimated correspond to the average impact, over our sample, of each variable.

Estimating the parameters in equation (7) and attributing them a causal interpretation requires addressing a number of empirical issues that relate to individual non-random spatial sorting and to various sources of endogeneity at both the individual and local levels.

The non-random spatial sorting issue relates to the fact that cities host workers with different individual abilities. For instance, the sorting of more-skilled workers into larger cities is observed in the US and in Europe. This can result either from the concentration of high-skill-intensive industries in larger cities or from the presence of consumption amenities (culture, night life, leisure) in these cities that attract more high-skilled workers. If this is also the case for China, then a positive correlation between average wages and city characteristics could simply reflect a composition effect, the over-representation of more-able workers in some cities. Standardly, such a bias can be addressed by using individual data to control for individual characteristics, that is, netting out individual wages from the role of s_i in equation (3), before assessing the role of city and city-industry characteristics. This is the strategy we adopt here through the vector X_i, which includes individual variables for gender, years of education, experience (and its square), self-employment (against salaried work), work in private sector (against public and collective sectors), and occupation (managerial work, technical work, office work, and service work). The unavailability of panel data covering all Chinese cities does not allow us to control for unobserved characteristics by using an individual fixed effect

as is done in some of the literature on developed countries[4]. However, Baum-Snow and Pavan (2012) argue that most of the spatial sorting in the US is explained by observed characteristics and that almost no sorting on unobserved characteristics occurs. Given further the small sorting on observed characteristics that we find in China (see below), sorting on unobservable characteristics uncorrelated to observed characteristics that would bias our estimation should not be very large. As a consequence, we do not see the absence of panel data as being particularly detrimental, although we acknowledge that sorting on unobservable characteristics should be further investigated for China when relevant data sets are available.

Most important, ordinary least squares estimates of specification (7) can be affected by various sources of endogeneity. As emphasised by Combes and Gobillon (2015), two issues are at stake. The first issue relates to endogeneity at the individual level, which results from individual location choices based on a precise job offer at wage w_i. As w_i includes the individual random component, this creates a correlation between the explanatory variables (which relate to the individual's actual location choice) and the residual. The only way to address that issue would be to use a natural random experiment, which would nevertheless pose a problem of external validity, or to use a structural approach as in Baum-Snow and Pavan (2012) for instance, an option not permitted by Chinese data. Crucially, in the Chinese context, where the *Hukou* system[5] has for years imposed strict restrictions on people's mobility, we do not see this as a critical source of endogeneity. The vast majority of urban natives has never migrated. This issue could be more severe for rural migrants but only if they were to receive individual job offers before migrating. This is highly unrealistic, as they are very low-skilled workers who take their migration decisions mostly based on expected returns only.

This last remark leads us to the second endogeneity issue, at the aggregate level, which is why we explicitly introduce a labour efficiency random aggregate component at the city-industry level, ν_{cs}^k, in the residual of specification (7). Indeed, the expected returns from migration, in particular the expected wage, depend not only on observed city and city-industry characteristics that are controlled for in the specification but also on ν_{cs}^k. This creates a spurious correlation between the explanatory variables and the residual when workers take their location decisions based on expected wages. This issue is potentially more problematic because migrants do take their location decision based on the expected wage, and it must be treated. The literature first proposes to isolate the individual and aggregate sources of endogeneity by estimating specification (7) not in one step but

[4]For further details on all estimation issues discussed in this section, see Combes and Gobillon (2015).

[5]The household registration -*Hukou*- system established in 1958 to monitor and limit population mobility assigns every Chinese citizen a registration status that records the place of residence and the 'agricultural' *versus* 'non-agricultural' status of the household. See Chan and Zhang (1999) for a discussion of the system.

in two steps, where city fixed effects are introduced in the first step instead of city variables, and next, the city fixed effects are explained in a second step by city variables that are instrumented. Because we want to address the possible endogeneity of both city and city-industry variables, we extend this strategy to three steps and consider the following specification:

$$\log w_i = X_i \theta^k + \delta_{cs}^k + \varepsilon_i^k, \tag{9}$$

$$\delta_{cs}^k = \textit{City-Industry variables}_{cs}\ \delta^k + \gamma_c^k + \nu_{cs}^k, \tag{10}$$

$$\gamma_c^k = \textit{City variables}_c\ \gamma^k + \xi_c^k, \qquad \text{for } k = H, L, M. \tag{11}$$

δ_{cs}^k is a group-k-specific city-industry fixed effect that captures the role of any city and city-industry characteristics, whether observed or not. As a result, OLS estimates of (9) could suffer from the non-random spatial sorting and individual endogeneity biases discussed above but not from aggregate endogeneity. Furthermore, the residual now encompasses individual random shocks only, which solves the possibly severe heteroscedasticity issues emphasised by Moulton (1990) when random shocks arise at both the individual and aggregate levels and some individuals move across locations.

Then, the second step (10) estimates how δ_{cs}^k relates to the vector of city-industry characteristics. These effects are estimated net of all observed and unobserved city characteristics that are controlled for through city fixed effects, γ_c^k. The last step (11) evaluates how city fixed effects depend on the city characteristics. ξ_c^k corresponds to the city-unexplained random component of wages. Both the second and third steps are possibly affected by aggregate endogeneity issues. Such endogeneity can emerge from reverse causality, e.g., some cities attract more firms and workers by offering higher expected returns, which in turn affects city characteristics. It can also result from certain city characteristics, in particular city endowments in productive public goods as universities or transport facilities for instance, that are missing in the vectors of city characteristics that we consider. The literature addresses both issues by using instrumental variables. Our three-step procedure that separates the role of city and city-industry characteristics partially alleviates the concern of an excessively large number of variables to be instrumented. The three city-industry characteristics that we consider are instrumented in the second step. As this estimation controls for city fixed effects, the instruments' exogeneity is easier to satisfy. However, finding relevant instruments at the city-industry level that allow for the identification of the effects in the within-city dimension remains demanding (which is possibly why we are, to the best of our knowledge, the first to attempt it). City characteristics are instrumented in the third step, which is more standard. Since seven city variables are introduced, we choose to instrument at most three of them simultaneously, as more than that

would be heroic in terms of identification power. We estimate these instrumented regressions either controlling for all or none of the non-instrumented variables and show that the results are consistent in both cases. The instruments and their corresponding identification assumptions are presented in the next section.

Finally, since the dependent variables in the second and third steps are estimated in the previous step, they may be plagued by measurement errors. We address this concern by using ordinary least squares weighted by the number of observations in the previous step. As the first-step estimation is weighted by the sampling weights provided in the survey we use to make it representative, the weight is the weighted number of observations. The logarithm of the variable is used for all variables in the regressions. For city-industry variables, their logarithm is then centered with respect to the city mean for each corresponding sample (high-skilled natives, low-skilled natives and rural migrants). As a consequence, these city means are part of the city fixed effects explained in the third step.

2.3 Data and variables

We use the 20% random extraction of the *1% 2005 China Population Survey* conducted by the National Bureau of Statistics of China. This allows us to define 257 cities akin to standard metropolitan areas and to disentangle 36 two-digit industries (listed in the supplementary material Table 1, Combes, Démurger, Li and Wang, 2019), each being represented in at least 33 and up to 227 cities. The survey is also comprehensive in including both registered urban residents and migrants (originating from all over China), which allows us to consider different groups of workers identified by their skill level and by their migration status. High-skilled workers are defined as workers who received at least technical or professional education after completing senior high school. Migrants are workers who have been working for at least 6 months in a county different from their county of *Hukou* registration, conditional on the two counties not being in the same core city. Depending on their type of registration ('agricultural' *versus* 'non-agricultural'), migrants are further grouped into rural *versus* urban migrants. We merge urban migrants with urban residents because they share similar observable characteristics, and we assume that rural migrants, high- and low-skilled, constitute a single specific input. Overall, we consider three groups of workers: high-skilled urban natives (composed of high-skilled urban residents and high-skilled urban migrants), low-skilled urban natives (composed of low-skilled urban residents and low-skilled urban migrants), and rural migrants (composed of both high-skilled and low-skilled rural migrants). We restrict the sample such that the three worker categories, high-skilled natives, low-skilled natives and rural migrants,

are present in any city-industry pair[6]. This leaves us with 245,935 non-agricultural workers located in 257 cities and operating in 36 industries, for a total of 2,554 city-industry pairs. High-skilled natives, low-skilled natives and rural migrants represent 20.6%, 52% and 27.4% of the overall sample of workers, respectively.

All city-sector and city level variables, except land area[7], are calculated based on the 20% subsample of the *1% 2005 China Population Survey* that we use for the first step of the estimation. A complete description of data sources and definition of the variables is provided in Appendix B. Summary statistics and the correlation between variables are displayed in the Supplementary material, Sections 2 and 3 (Combes et al., 2019).

Our instrumentation strategy for both city-industry and city variables first considers a number of historical instruments, as commonly used in the literature (Combes and Gobillon, 2015). The intuition is that current productivity shocks should not be correlated with the employment structure decades before the date of observation (here, 2005). This is probably a particularly accurate assumption for China, which experienced a transition from a fully planned, closed economy to a market-based economy largely open to international trade. Since we move as far in time as 1964, considerable political changes also likely generally disconnect past variables from current shocks. On the other hand, historical instruments remain, in general, relevant because there is considerable inertia in the urban hierarchy, which we expect to be driven by political and cultural factors that, again, do not relate to current economic conditions. Specifically, instruments for both city and city-industry variables are extracted or computed from the *Historical China County Population Census Data* for 1964, 1982, and 1990, which is provided by the University of Michigan China Data Centre. One concern in using historical data is that city boundaries have changed over time. Hence, when computing historical instruments, we use the core city boundaries if the city was a prefecture-level city at the census date, the county-level city boundaries if the city was not a prefecture-level city then, and the county boundaries if it was not even a county-level city. These boundaries do not match current boundaries, but they remain correlated with them, thus favouring both exogeneity and relevance.

We also construct specific instruments for the city-industry and city migrant ratios, S_{cs}^{M} and S_{c}^{M}, which are *a priori* the most affected by reverse causality. We follow the approach initiated by

[6]Hence, worker-category-specific estimates are based on the same city-industry observations. We verified that estimations on the full unbalanced sample lead to similar conclusions.

[7]City land area comes from the *China City Statistical Yearbook* and corresponds to the area of the core city. As detailed in Appendix B, it is larger than the urban parts of the core city that correspond to our definition of the metropolitan area. As such, this creates sources of measurement error, although they should not be particularly large or have substantial adverse effects on our results.

Altonji and Card (1991) and Card (2001) who use lagged and predicted migrant flows, respectively [8]. Again, this assumes that migration patterns are relatively stable over time, which makes past flows relevant. Simultaneously, as expected wages are rapidly affected by economic changes, current wage shocks not related to our observed city characteristics should be largely disconnected from historical flows. On the other hand, reverse causality would not be eliminated by historical instruments if large historical migrant flows had made some cities anticipate large future flows and invest more in anticipation of a larger labour pool.

We first use the number of rural migrants in 1990, and we believe that this measure should be free of most of the endogeneity bias. Indeed, although it began in the early 1980s, internal migration in China did not reach dramatically rapid growth until the mid-1990s, and the year 1990 can realistically be considered a pre-reform year in that respect. The average city share of migrants in the population was fairly small at that date, at 1.9%, and it increased by nearly a factor of 10 to reach 18% in 2015. Moreover, many changes that accelerated China's movement towards a market economy occurred from the mid-1990s only (including the accession to the WTO in 2001). Finally, most city characteristics that could drive migration choices are controlled for, including city size, industrial composition, and market access.

Our second strategy uses predictors of migrant flows. It consists in computing, for each Chinese province (administrative units larger than cities), the share of its emigrants to any destination city and then multiplying this share by a proxy for the total number of the province's emigrants. The predicted number of migrants in a city is obtained by summing over all origin provinces. Baum-Snow, Brandt, Henderson, Turner and Zhang (2017) use the actual provinces' share of migrants over 1985–1990 and multiply it by the actual number of migrants over 1995–2000 to instrument their 2010 variables. As suggested by Meng and Zhang (2010), further exogeneity can be gained by using predicted, and not actual, provincial migrant shares and total number of emigrants. Migrant share predictions are obtained by first estimating a gravity model, where 1995–2000 flows are explained by province (origin) and city (destination) fixed effects, bilateral distance and two dummy variables for the city being either within the origin province or in a contiguous province (to capture so-called 'border' effects). Meng and Zhang (2010) use push factors (land, income and physical investment per capita, and total land area subject to natural disasters in the province) to predict the province's total number of emigrants, and they use the 2000–2005 time variation in the predicted number of migrants in the city to instrument the 2005 level. We predict more directly the province's total number of emigrants by using the province's 2000 total employment, which should add further

[8] See Lewis and Peri (2015) for a survey of the relevant literature.

exogeneity. We obtain a predicted number of migrants in the city by summing over all provinces the product of the predicted share of emigrants to the city by the province's total employment. For either strategy, the ratio of the historical (predicted) number of migrants over the number of low-skilled workers or over total employment at corresponding dates, 1990 and 2000, is used as the actual instrument.

The advantage of the first strategy comes from longer time lags, but it uses the actual number of city migrants, while the second method uses only predicted migrant shares and the predicted total number of emigrants at the province level with, however, shorter lags[9]. We implement similar strategies at the industry level to construct predictors of city-industry migrants. Appendix B provides further details on how migrant instruments are computed.

The instrumentation of the three city-industry characteristics in the second step is the most demanding one. We use as instruments the 1990 values of the three city-industry characteristics that enter the specification, and we add the share of predicted 2000 city-industry migrants in the 2000 city population, computed as explained above. For the sample of migrant workers, we also use the 1995 share of state-owned firms' employment in total employment as an additional instrument that allows us to further capture how high-skilled, low-skilled and migrant workers split across city-industries. Overall, these city-industry instruments are somewhat weak (recall that we control for city fixed effects), and some effects lose significance compared to OLS, which makes us view these estimations as a first assessment that will deserve further attention in the future. On the other hand, we show that some interesting conclusions at the city-industry level seem to be robust to instrumentation. The complete list of instruments used for city-industry characteristics is provided in Appendix D, Table 6, Panel (a).

Regarding city characteristics, we instrument density, land area and the rural to low-skilled workers ratio, the other variables being introduced as controls only. We have a larger set of possible instruments, and we experimented with many combinations of them, all of which yielded largely consistent results with one another and with the OLS. We choose to be parsimonious, and we report estimations for different groups of workers using the same sets of instruments to allow for reliable comparisons. Overall, in addition to the migrant instrument, we use two groups of historical variables only. The first group consists of employment density in 1964, land area in 1982 and the share of agricultural employment in 1982. While the migrant instrument should capture variations in the migrant over low-skilled workers ratio, the other three instrumental variables can clearly capture variations in current density and land area, with the share of agriculture being a further predictor

[9]Variants of these instruments, considering for instance total rural employment in the province to predict the total number of emigrants, lead to similar conclusions.

of the divide among workers in terms of high-skilled, low-skilled, and rural migrant workers. The second group of historical instruments is constructed in the same spirit and includes the share of non-agricultural employment in 1964, land area in 1982, and the share of college students in 1982, again helping to capture the employment composition in terms of skills. In the main text, we complement the historical variables by the predicted number of migrants in 2000 over the city's year 2000 total employment, while in robustness checks in the appendix, we complement them by the 1990 number of migrants over the 1990 number of low-skilled workers. The list of instruments used for density, land area and the rural to low-skilled workers ratio is displayed in Appendix D, Table 6, Panels (b) and (c).

3 Results

3.1 Location matters

Table 1 presents a variance analysis for the three steps of the estimation, in the vein of Abowd, Kramarz and Margolis (1999). It uses the estimated OLS parameters reported in Appendix C, Table 5 for the first (individual) step, and in Table 2, columns (1), (4) and (7) for the second (city-industry) and the third (city) steps[10]. It reads as follows. The relative explanatory power of any variable is large when its effect both varies substantially relative to the variations in the dependent variable (its standard variation reported in column 'St.D.' is large) and is highly correlated with it (column 'Corr. 1' is large). Moreover, the magnitude of spatial sorting –or the extent to which workers with better characteristics tend to locate in city-industries with higher returns– is evaluated through the correlation between the effect of individual variables and the city-industry fixed effects ('Sort').

The first striking result from Table 1 relates to the large explanatory power of city-industry fixed effects, which is higher than the explanatory power of individual characteristics[11]. This result holds for all three groups of workers and most prominently for high-skilled workers, in sharp contrast with findings from more developed countries (the US, France, or the UK, for instance; see Baum-Snow and Pavan (2012), Combes, Duranton and Gobillon (2008) and D'Costa and Overman (2014),

[10]For a simplified model that would encompass two explanatory variables only, x_i and z_i, and city-industry fixed effects δ_{cs}, $y_i = \alpha x_i + \gamma z_i + \delta_{cs} + \varepsilon_i$, the first line in Table 1 reports the standard deviation of the explained variable, y_i, in column 'St.D.'. The second line of the same column reports the standard deviation of $\alpha\, x_i$, that is, the standard deviation of the effect of x_i on y_i everything else equal. On the same second line, column 'Corr. 1' displays the correlation between y_i and $\alpha\, x_i$. Furthermore, for the second and third steps (Panels (b) and (c)), the correlation of the effect of the variables can be computed either with the individual wage ('Corr. 1') or with their own dependent variable,'Corr. 2' and 'Corr. 3'.

[11]The first-step estimation results reported in Appendix C highlight standard findings on the marginal impact of the usual Mincerian individual variables (gender, education, experience, enterprise ownership and occupation).

Table 1 – Variance analysis

	High-Skilled urban natives			Low-Skilled urban natives			Rural migrants		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel (a): First step	St.D.	Corr1	Sort	St.D.	Corr.1	Sort.	St.D.	Corr.1	Sort.
Dependent: Individual wage	0.673	1.00		0.654	1.00		0.552	1.00	
All individual charac.	0.229	0.40	0.10	0.233	0.40	0.10	0.204	0.37	0.01
Male	0.0459	0.11	0.01	0.101	0.20	0.06	0.0644	0.19	0.05
Education	0.144	0.32	0.15	0.110	0.25	0.07	0.133	0.29	0.06
Experience	0.114	0.13	−0.10	0.0480	0.11	−0.02	0.0433	0.06	−0.01
Occupation	0.0935	0.28	0.13	0.109	0.26	0.11	0.0953	0.21	−0.05
Self-employed	0.0142	−0.02	−0.09	0.0662	−0.03	−0.17	0.0346	0.03	−0.15
Private firm	0.0153	−0.01	0.07	0.072	0.20	0.14	0.0222	0.07	0.03
City-industry fixed effects	0.397	0.62	1.00	0.304	0.50	1.00	0.240	0.44	1.00
First-step residual	0.473	0.70	0.00	0.517	0.79	0.00	0.453	0.82	0.00
Panel (b): Second step	St.D.	Corr.1	Corr.2	St.D.	Corr.1	Corr.2	St.D.	Corr.1	Corr.2
All city-industry charac.	0.0418	0.14	0.12	0.0418	0.10	0.10	0.0260	0.12	0.11
Migrant/Low$_{cs}$	0.0005	0.08	0.05	0.0157	0.11	0.13	0.0187	0.09	0.07
High/Low$_{cs}$	0.0418	0.15	0.14	0.0317	0.15	0.15	0.0178	0.11	0.14
Specialisation$_{cs}$	0.0080	−0.05	−0.07	0.0274	−0.08	−0.09	0.0101	−0.06	−0.08
Industry fixed effects	0.079	0.20	0.31	0.0992	0.22	0.33	0.0927	0.19	0.38
City fixed effects	0.345	0.55	0.89	0.256	0.40	0.85	0.166	0.28	0.69
Second-step residuals	0.155	0.23	0.39	0.124	0.20	0.41	0.146	0.26	0.61
Panel (c): Third step	St.D.	Corr.1	Corr.3	St.D.	Corr.1	Corr.3	St.D.	Corr.1	Corr.3
All city charac.	0.317	0.51	0.92	0.226	0.36	0.88	0.133	0.22	0.81
Migrant/Low$_{c}$	0.197	0.47	0.86	0.134	0.33	0.81	0.0628	0.14	0.58
Density$_{c}$	0.0531	0.23	0.43	0.0556	0.18	0.46	0.0257	0.12	0.47
Land area$_{c}$	0.106	0.27	0.45	0.0679	0.16	0.31	0.0149	0.11	0.29
High/Low$_{c}$	0.0295	0.27	0.43	0.0138	0.18	0.26	0.0223	0.14	0.36
Market access$_{c}$	0.0547	0.21	0.41	0.0638	0.18	0.51	0.0283	0.10	0.45
Port proximity$_{c}$	0.0439	0.28	0.52	0.0361	0.22	0.53	0.0500	0.16	0.57
Diversity$_{c}$	0.0146	−0.18	−0.31	0.0347	−0.15	−0.30	0.0253	0.08	0.20
Third-step residuals	0.134	0.20	0.39	0.119	0.17	0.47	0.0979	0.17	0.59

Notes: 'St.D.': Standard deviation of the effect, ' Corr.1': Correlation with individual wage (first-step dependent), 'Sort.': Sorting, correlation with the city-industry fixed effect, 'Corr.2': Correlation with the city-industry fixed effect (second-step dependent), 'Corr.3': Correlation with city fixed effect (third-step dependent). Line 'All individual charac.' corresponds to the total effect of all individual characteristics, with individual effects detailed in the indented lines just below. Estimations are based on 50,834 observations for high-skilled workers, 127,750 observations for low-skilled workers and 67,351 observations for rural migrants. Individual-level parameters are estimated in Appendix C, Table 5. Line 'All city-industry charac.' corresponds to the total effect of all city-industry characteristics, with individual effects detailed in the indented lines just below. City-industry-level parameters are estimated in Table 2, columns (1), (4) and (7). Line 'All city charac.' corresponds to the total effect of all city characteristics, with individual effects detailed in the indented lines just below. City-level parameters are estimated in Table 2, columns (1), (4) and (7). The variables are defined in Appendix B.

respectively), where the explanatory power of individual characteristics is found to be larger than that of local effects. Therefore, if the fact that location matters for individual wages is now well documented for developed countries, it seems that it matters even more in China. Workers with identical individual characteristics would earn more in some city-industry cells than in others.

Conversely, the literature on developed countries also finds evidence of spatial sorting: workers with better individual characteristics tend to concentrate in more-rewarding locations. The second striking feature from Table 1 is that spatial sorting does not seem to occur in China or at best occurs only to a fairly small extent, mostly for urban natives, and not for migrants (see column 'Sort.'). The joint effect of all individual characteristics is correlated with city-industry fixed effects at 0.10 for either high- or low-skilled natives and at 0.01 for migrants only[12]. Sorting according to some characteristics is slightly higher (e.g., education for high-skilled workers, self-employment or private firm employment for low-skilled workers), but positive sorting and negative sorting compensate for one another and lead to this small positive overall sorting.

3.2 City-industry migration and urbanisation gains

Table 2 Panel (a) reports OLS and IV estimates for the second step, on the determinants of city-industry fixed effects. As detailed in Section 2.1, these determinants include both city-industry characteristics and city fixed effects. As a complement, Table 7 in Appendix E illustrates the magnitude of each city-industry and city effect, computed as the percentage increase in wages (reported in column '%') when moving from the first quartile to the last quartile of the city-industry or city variable (the inter-quartile difference, 'P75-P25' is provided in the first column)[13]. Following Stock and Yogo (2005), we use the limited information maximum likelihood (LIML) estimator for instrumented regressions because our instruments are not very strong, particularly at the city-industry level. The Cragg-Donald statistics obtained are all above 5, that is, above the 5% critical values for LIML given the number of instruments and instrumented variables we consider.

As displayed in Table 2, Panel (a), there is evidence of substitution effects within industries in Chinese cities for migrants and to a lesser degree for low-skilled native workers. Consistent with the intuition, migrants incur losses from the presence of other migrants within an industry. The magnitude of the effect when the ratio of migrants over low-skilled migrants moves from the first to the last quartile is not very large, however, as it ranges from -2.2% in OLS estimation (Table 7, Appendix E) to -4.6% in instrumented estimations, and the large confidence interval

[12]For comparison, Combes et al. (2008) highlight a larger correlation for France, at 0.29 (although this number also accounts for the role of individual fixed effects).

[13]The computations are based on OLS estimates that include all the city-industry and city variables (columns (1), (4), and (7), Table 2).

suggests that the difference between the estimations is not significant. Low-skilled urban natives also seem to be negatively affected by the ratio of migrants over low-skilled workers but to a lesser extent. OLS estimates give a magnitude of the effect of -2%, but the effect becomes non-significant when city-industry variables are instrumented. These findings indicate that the substitution effect may dominate within an industry: migrants seem to be fairly substitutable for low-skilled workers, and the externality effect, if any, is not large enough to compensate for this. However, the small economic magnitude of the effect suggests that low-skilled workers do not loose much from the presence of migrants in their city-industry. At the other end of the skill distribution, high-skilled urban natives do not appear to be impacted, neither positively nor negatively, by the presence of migrants within their industry.

Some evidence of human capital externality is also found at the industry level for all three groups of workers, although the effect is again small and not fully robust. OLS estimates suggest that high-skilled workers benefit from human capital externalities at the city-industry level, with the externality effect dominating the substitution effect. Moving from the first quartile to the last quartile of the city-industry high-skilled ratio, the associated high-skilled workers' wage would increase in the city-industry by 7.2%. However, instrumented regressions indicate that this effect could be non-significant, and we will see below that instrumented estimations also challenge the significance of human capital effect at the city level. Similarly for migrants, OLS estimates would conclude that some human capital externalities are present but to a small extent (2.1% for the inter-quartile), and the effect becomes non-significant when instrumented. On the other hand, low-skilled urban natives seem to benefit more from city externalities within their industry, arising both from the presence of high-skilled workers and from specialisation, a conclusion that matches previous findings on other countries (see Combes and Gobillon, 2015). Regarding human capital externalities, low-skilled workers' wages are larger by 3.5%, and up to 18.1% (with a fairly large confidence interval, however) according to instrumented regressions, when the ratio of high-skilled over low-skilled workers moves from the first to the last quartile. Low-skilled urban natives are also the only group that benefits from specialisation. The magnitude of the impact is similar to what is found in developed countries, at 3.8% for the inter-quartile gap, and even slightly more if one uses instrumented estimations.

Finally, and as a complement to our estimates, the variance analysis reported in Table 1 Panel (b) suggests that city-industry characteristics considered together explain little of the individual earnings disparities for all three categories of workers[14]. Conversely, city fixed effects explain most

[14]Combes et al. (2008) report a similar finding for France.

of city-industry fixed effects, substantially so for high-skilled workers, slightly less so for low-skilled workers and considerably less so for migrants, indicating again that location matters more for urban (especially high-skilled) workers than for migrant workers. It is therefore important to assess the relative importance of overall city characteristics, which we do in the next section.

Table 2 – OLS and IV estimates for Steps 2 and 3

	High-skilled urban natives			Low-skilled urban natives			Rural migrants		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel (a): Second st.	OLS	IV		OLS	IV		OLS	IV	
Migrant/Low$_{cs}$	-0.000480	-0.0615		-0.0177^{a}	0.0106		-0.0254^{a}	-0.0535^{b}	
	(0.00491)	(0.0393)		(0.00377)	(0.0205)		(0.00603)	(0.0216)	
High/Low$_{cs}$	0.0400^{a}	-0.0230		0.0347^{a}	0.167^{a}		0.0231^{a}	-0.0393	
	(0.00795)	(0.0690)		(0.00507)	(0.0435)		(0.00704)	(0.0600)	
Specialisation$_{cs}$	0.0114	-0.00395		0.0357^{a}	0.0428^{a}		0.0127^{b}	-0.0293	
	(0.00716)	(0.0222)		(0.00523)	(0.0165)		(0.00604)	(0.0163)	
Observations	2,554	2,554		2,554	2,554		2,554	2,554	
R2	0.85			0.83			0.63		
Weakness		5.2			8.7			5.8	
Over-identification		0.22			0.36			0.64	
Panel (b): Third st.	OLS	IV1	IV2	OLS	IV1	IV2	OLS	IV1	IV2
Migrant/Low$_{c}$	0.193^{a}	0.241^{a}	0.204^{a}	0.123^{a}	0.139^{a}	0.116^{a}	0.0528^{a}	0.0500^{a}	0.0534^{a}
	(0.0131)	(0.0201)	(0.0239)	(0.0101)	(0.0165)	(0.0162)	(0.00977)	(0.0117)	(0.0151)
Density$_{c}$	0.0658^{a}	0.141^{a}	0.159^{a}	0.0643^{a}	0.132^{a}	0.0701^{b}	0.0284^{b}	0.0445^{b}	-0.0153
	(0.0180)	(0.0333)	(0.0558)	(0.0139)	(0.0241)	(0.0282)	(0.0116)	(0.0196)	(0.0213)
Land area$_{c}$	0.109^{a}	0.0839^{a}	0.131^{a}	0.0695^{a}	0.0595^{a}	0.0851^{a}	0.0183	0.0787^{a}	-0.0197
	(0.0156)	(0.0201)	(0.0422)	(0.0127)	(0.0172)	(0.0242)	(0.0127)	(0.0153)	(0.0347)
High/Low$_{c}$	0.0624^{a}		0.00310	0.0235		0.0245	0.0394^{b}		0.0821^{a}
	(0.0238)		(0.0382)	(0.0170)		(0.0191)	(0.0158)		(0.0210)
Market access$_{c}$	0.192^{a}		0.0865	0.211^{a}		0.225^{a}	0.104^{a}		0.131^{a}
	(0.0382)		(0.0506)	(0.0322)		(0.0428)	(0.0306)		(0.0375)
Port proximity$_{c}$	0.0218^{a}		0.0131	0.0187^{a}		0.0178^{a}	0.0270^{a}		0.0360^{a}
	(0.00566)		(0.00820)	(0.00497)		(0.00549)	(0.00466)		(0.00622)
Diversity$_{c}$	-0.0711		-0.0771	-0.142^{a}		-0.167^{a}	0.101^{b}		0.0854
	(0.0575)		(0.0908)	(0.0449)		(0.0544)	(0.0403)		(0.0649)
Observations	257	257	257	257	257	257	257	257	257
R2	0.85			0.78			0.65		
Weakness		21.3	8.3		19.6	10.0		24.5	9.85
Over-identification		0.54	0.42		0.69	0.48		0.21	0.09

Notes: Instrumented estimations performed with LIML. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B. Panel (a): The dependent variable is the city-industry fixed effect estimated in step 1. All columns include industry fixed effects. In columns labeled 'IV', all three variables are instrumented. Panel (b): The dependent variable is the city fixed effect estimated in step 2. In columns labeled 'IV1' and 'IV2', density, land area and the rural to low-skilled workers ratio are instrumented. Other variables are introduced as controls only in columns labeled 'IV2'. Instruments at the city-industry level and city level are defined in Appendix B, and those used for each column are listed in Table 6, Appendix D. The line labeled 'over-identification' reports the p-value for the Hansen J-test; the line labeled 'weakness' reports the Cragg-Donald statistics. Following Stock and Yogo (2005), critical values for LIML estimates at the 5% level are below 5 for all estimations reported in the table.

3.3 Migration and urbanisation gains at the city level

In contrast with the second step, the variance analysis reported in Table 1, Panel (c) shows that, at the city level, city characteristics explain the city fixed effects very well. The migrant ratio to low-skilled workers has the largest explanatory power, by very far for high-skilled, by far for low-skilled, and somewhat for migrant workers. Apart from the migrant effect, city density, land area, and market access (both internal and international) matter the most for high- and low-skilled workers, while for migrants, the second most important variable is the proximity to a seaport, which is consistent with the fact that the export sector is fed by inflows of migrants. Finally, and consistent with the urban economics literature (see Combes and Gobillon, 2015), diversity has the lowest explanatory power, and its marginal effect, often non-significant, is the least robust.

Consistent with the variance analysis and our focus on migration and urbanisation gains, we present IV estimates for the impact of the migrant share, density and land area, with and without the other city variables as controls. As noted in Section 2.2, instrumenting more city variables would be too demanding. The results are presented in Table 2, Panel (b) with historical city characteristics and the ratio of the predicted number of migrants over total employment in 2000 as instruments. A second set of instrumented estimations, replacing the migrant instrument with the 1990 ratio of migrants over low-skilled workers, is provided in Table 8, Appendix F.

The most striking finding from Table 2 relates to the large positive migrant impact at the city level, for both high- and low-skilled urban natives. Although to a lesser extent, migrant workers also seem to benefit from their own presence. Moreover, the role of migrants does not hinder the simultaneous presence of agglomeration gains from both density and land area, which are similarly fairly large for urban natives and smaller for migrants. Further gains also arise from market access and proximity to seaports.

Before detailing the magnitude of the effects, another general comment worth emphasising is that instrumentation that uses the year 2000 predicted number of migrants tends to magnify the impact of density, land area and migrants for high-skilled workers, while it reduces the role of human capital externalities and market potential. For low-skilled workers, the OLS and IV estimations lead to very similar marginal effects for all variables. For migrants, the relatively more limited roles of migrants and city size appear to be more difficult to disentangle from one another. This is not entirely a surprise given the positive correlation between migrant variables and most of the other city characteristics for this group of workers. As reported in Combes et al. (2019), Table 14, the correlation of the migrant ratio with density is 0.63 for instance. Instrumented estimations that use predicted year 2000 migrants tend to maintain the positive effect of migrants on themselves

while eliminating the impact of density and land area. By contrast, instrumented estimations that use the 1990 number of migrants (Table 8, Appendix F) lead to the opposite finding: the migrant effect vanishes, while the density and land area impacts are not only confirmed but become as high as those for the low-skilled workers. Using other historical and migrant instruments (not reported here) confirms the conclusion that one or the other polar case emerges. The second instrument also leads to slightly smaller magnitudes of the migrant impact for high- and low-skilled urban natives.

Generally speaking, not controlling for other city characteristics (market access, proximity to seaports, high-skilled ratio and diversity) would lead us to conclude that migrants, density and land area have even larger impacts. This is well documented for other countries, and it mostly results from the positive correlation of density with market access variables (Combes et al., 2019). Even if instrumented estimations should partly address the issue, we prefer estimations that control for other city characteristics. However, given that doing so comes at the risk of introducing endogenous controls, we check that the estimations without controls do not lead to discordant results.

We now turn to the magnitude of the estimated effects for the different groups of workers. High-skilled workers gain the most from migration and urbanisation. As shown in Table 7 of Appendix E, moving from the first to the last quartile of the migrant variable increases high-skilled workers' wages by 27.7%. High-skilled workers also gain from density, with wages being 6.9% larger for the inter-quartile range, and from cities larger in terms of land area, by 16.4% for the inter-quartile range. Even when one takes the smallest impact of migrants from the IV estimates, the gain from migrants is 21.8%, with a density gain now at 15.8% and a gain from land area at 14.3%. From an international perspective, these numbers are large: the density impact is above the highest levels obtained for high-skilled workers in developed countries; land area is usually not found to provide any extra gain; and an effect comparable to that of rural migrants has never been documented for other countries. According to both OLS estimates and the instrumented regressions using the 1990 number of migrants (Table 8, Appendix F), high-skilled workers further gain from human capital externalities, market access and proximity to seaports. These are however, second-order effects, as their magnitudes are lower (at most 4.7%, 7.0% and 5.0%, respectively, for the inter-quartile range) and the impact is not robust to instrumentation with the predicted year 2000 migrants, as shown in Table 2.

Low-skilled workers also gain from migration and urbanisation. All city characteristics that have a significant impact on high-skilled workers' wages are also significant for low-skilled workers. However, their impact is not as large, with the exception of market access and proximity to seaports, the impacts of which remain significant in the instrumented regressions. Typically, for low-skilled

workers, the gain from migrants at the city level is 23.8% for the inter-quartile range according to the OLS estimates and not lower than 14.5% in the instrumented regressions. Similarly, the gain from density ranges between 7.8% and 8.5% and the gain from land area between 7.0% and 10.2%. Additional gains of 8.2% up to 10.2% from market access and of 3.4% up to 4.6% from proximity to a seaport are also observed. As generally found in the urban economics literature, diversity has no significant effect, except potentially a small negative impact for low-skilled natives.

Finally, rural migrants who generate large gains for urban natives also benefit from the presence of migrants and from agglomeration, although their gain is substantially lower in magnitude than that of urban natives. Gains from migrants in the city are 7.7% for the inter-quartile range (both OLS and IV estimates in Table 2), in addition to which there are gains of 3.1% and 1.6% from density and land area. Migrants also benefit from market access (3.1%) and proximity to a seaport (4.0%). As discussed above, the estimation using instruments based on the 1990 number of migrants (Table 8) indicates larger gains from density (8.2%), land area (8.6%), and market access (7.1%) and gains from human capital externalities (up to 8.3%), but non-significant gains from migrants. Hence, the results with respect to which city characteristics matter the most are slightly less uniform for migrants. Nevertheless, the overall conclusion that they benefit from urbanisation (but to a lesser extent than urban natives) is clear.

How do city characteristics jointly affect wage inequality among the different groups of workers? Table 3 reports a Oaxaca-Blinder decomposition exercise that provides some answers. It is based on OLS estimates displayed in Table 2. In particular, this ignores a possible heterogeneity of estimated parameters across cities and industries yet suggested by theory (see equation 6) and some other possible endogeneity issues. The columns 'Composition' and 'Return' report the contribution of the composition effect and of differences in returns, respectively, for all individual variables and the city characteristics.

First, the raw wage gap is unsurprisingly high between high-skilled workers and both low-skilled workers (56.6%) and migrants (66.6%), with a 10 percentage point difference in the gap that highlights the disadvantaged position of migrants, even compared to low-skilled urban natives. Second, and consistent with findings widely documented in the literature (see, for instance, Démurger, Gurgand, Li and Yue, 2009), wage inequality between the different groups of workers in China is in part explained by their different individual characteristics. Both better individual characteristics and higher returns from these characteristics account for an important share of the wage gap between high-skilled workers and migrants (16.6% and 46.3% of the gap, respectively). A similar pattern holds for the difference between high-skilled and low-skilled workers and between low-skilled and

migrant workers, with yet lower gaps.

Third, we document here that another large share of the wage gaps is driven by different returns to urbanisation and operating in different cities. Indeed, the city in which workers locate is crucial. Interestingly, migrants are locating in cities that are more rewarding, and this location choice contributes to partly compensating for the fact that city returns are lower for them compared with both high-skilled and low-skilled workers: While 112.6% (resp. 93.4%) of the gap between high-skilled (resp. low-skilled) workers and migrants is because the former have better returns to city characteristics, 9.1% (resp. 14.3%) of the gap is, by contrast, compensated by an average location of migrants in better cities (e.g., denser, with more migrants, or with better access to markets).

Table 3 – Oaxaca-Blinder Decomposition

	High vs Migrants (gap: 66.6%)		Low vs Migrants (gap: 6.4%)		High vs Low (gap: 56.6%)	
	Composition	Return	Composition	Return	Composition	Return
All individual charac.	16.6%	46.3%	9.9%	−2.4%	3.5%	53.9%
All city charac.	−9.1%	112.6%	−14.3%	93.4%	6.1%	10.9%

Notes: Oaxaca-Blinder decomposition based on OLS estimates displayed in Table 2. Note that the city-industry characteristics do not contribute to the wage gap by construction because they are centered with respect to the city mean for the corresponding sample (high-skilled natives, low-skilled natives and rural migrants).

4 Conclusion

Location matters for explaining individual earnings in urban China, but it matters to different extents for various groups of workers. The different returns to urbanisation contribute to the role of different individual characteristics that shape wage gaps in urban China. High-skilled urban workers earn considerably more than migrant workers – and more than low-skilled workers – not only because they have better individual characteristics and derive higher returns from these characteristics, but also because they benefit more from urbanisation through higher returns, despite the fact that they tend to locate in less-rewarding locations compared to migrants.

We also document that beyond standard urban features that relate to city size and the proximity to large markets, both within the country and at the international level, rural migrants play a particular role in urbanisation effects in China. Far from inducing losses for urban natives, they generate fairly considerable gains and contribute substantially to increasing labour productivity in Chinese cities. However, rural migrants themselves face strong competition, which is partly reflected in the negative substitution effect we estimate at the industry-city level for rural migrants, and more

generally, in their lower benefit from urbanisation.

A difficulty in assessing the impact of city-industry and city characteristics relates to the possible reverse causality between these variables and earnings. We exercise considerable caution in addressing this issue by using a large set of instruments different in nature and by recognising the fact that Chinese cities were affected by large shocks in past decades and thus computing the instruments before these shocks. Our findings remain to be further corroborated with different sources of data and estimation strategies. In the future, researchers should benefit from extended data sets that could make it possible to control for unobserved individual effects, even if at present individual spatial sorting does not seem to be particularly large. Endogenous location choices could change that in the future. The city characteristics that are the most important for workers' earnings today could also change as a result of the ongoing urbanisation that continuously reshapes them.

Finally, the impact of city and city-industry characteristics is interpreted as shaping individual productivity, which, in turn, affects nominal earnings. The extent to which workers' real earnings, and thus individual utility, are affected by local characteristics and how this varies across groups of workers would require having access to city cost of living measures and to endowments in local amenities. Although not feasible at present due to a lack of data, such extensions would further refine the assessment of the impact of urbanisation on individual inequality by evaluating the extent to which nominal wage gaps are dulled by city costs of living. It would also possibly allow for policies to enhance both spatial efficiency and individual equity. Further understanding the role that cities play in both productivity and real earnings in China thus remains high on the research agenda.

References

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999), 'High wage workers and high wage firms', *Econometrica* **67**(2), 251–333.

Albouy, D. (2008), Are big cities really bad places to live? Improving quality-of-life estimates across cities, Working Paper 14472, National Bureau of Economic Research.

Altonji, J. G. and Card, D. (1991), The effects of immigration on the labor market outcomes of less-skilled natives, *in* J. Abowd and R. B. Freeman, eds, 'Immigration, trade, and the labor market', University of Chicago Press, pp. 201–234.

Au, C. and Henderson, V. (2006), 'Are Chinese cities too small?', *Review of Economic Studies* **73**, 549–576.

Baum-Snow, N., Brandt, L., Henderson, V. J., Turner, M. A. and Zhang, Q. (2017), 'Roads, railroads and decentralization of chinese cities', *Review of Economics and Statistics* **99**(3).

Baum-Snow, N. and Pavan, R. (2012), 'Understanding the city size wage gap', *Review of Economic Studies* **79**(1), 88–127.

Card, D. (2001), 'Immigrant inflows, native outflows, and the local labor market impacts of higher immigration', *Journal of Labor Economics* **19**(1), 22–64.

Chan, K. W. and Hu, Y. (2003), 'Urbanization in China in the 1990s: New definition, different series, and revised trends', *The China Review* **3**, 49–71.

Chan, K. W. and Zhang, L. (1999), 'The Hukou system and rural-urban migration in China: Processes and changes', *China Quarterly* **160**, 818–855.

Chauvin, J. P., Glaeser, E., Ma, Y. and Tobio, K. (2017), 'What is different about urbanization in rich and poor countries? Cities in Brazil, China, India and the United States', *Journal of Urban Economics* **98**, 17–49.

Combes, P.-P., Démurger, S. and Li, S. (2015), 'Migration externalities in China', *European Economic Review* (76), 152–167.

Combes, P.-P., Démurger, S., Li, S. and Wang, J. (2019), Supplementary material for 'Unequal migration and urbanisation gains in China', Available at www.xxx.

Combes, P.-P., Duranton, G. and Gobillon, L. (2008), 'Spatial wage disparities: Sorting matters!', *Journal of Urban Economics* **63**(2), 723–742.

Combes, P.-P. and Gobillon, L. (2015), The empirics of agglomeration economies, *in* G. Duranton, V. Henderson and W. Strange, eds, 'Handbook of Urban and Regional Economics', Vol. 5A, North-Holland, Amsterdam.

D'Costa, S. and Overman, H. (2014), 'The urban wage growth premium: Sorting or learning?', *Regional Science and Urban Economics* **48**, 168–179.

de Sousa, J. and Poncet, S. (2011), 'How are wages set in Beijing?', *Regional Science and Urban Economics* **41**, 9–19.

Démurger, S., Gurgand, M., Li, S. and Yue, X. (2009), 'Migrants as second-class workers in urban China? A decomposition analysis', *Journal of Comparative Economics* **37**(4), 610–628.

Han, J. and Li, S. (2017), 'Internal migration and external benefit: The impact of labor migration on the wage structure in urban China', *China Economic Review* **46**, 67–86.

Harris, C. (1954), 'The market as a factor in the localization of industry in the United States', *Annals of the Association of American Geographers* **44**(4), 315–?348.

Head, K. and Mayer, T. (2004), The empirics of agglomeration and trade, *in* V. Henderson and J.-F. Thisse, eds, 'Handbook of Regional and Urban Economics', Vol. 4, North-Holland, Amsterdam, pp. 2609–2669.

Hering, L. and Poncet, S. (2010), 'Market access and individual wages: Evidence from China', *Review of Economics and Statistics,* **92**, 145–159.

Jacobs, J. (1969), *The Economy of Cities*, Random House, New York.

Lewis, E. and Peri, G. (2015), Immigration and the economy of cities and regions, *in* G. Duranton, V. Henderson and W. Strange, eds, 'Handbook of Urban and Regional Economics', Vol. 5A, North-Holland, Amsterdam.

Li, S. and Sicular, T. (2014), 'The distribution of household income in china: Inequality, poverty and policies', *The China Quarterly* **217**, 1–41.

Liu, Z. (2014), 'Human capital externalities in cities: Evidence from Chinese manufacturing firms', *Journal of Economic Geography* **14**(3), 621–649.

Marshall, A. (1890), *Principles of Economics*, Macmillan, London.

Meng, X. and Zhang, D. (2010), 'Labour market impact of large scale internal migration on Chinese urban 'native' workers', *IZA Discussion Paper* **5288**.

Moretti, E. (2004), Human capital externalities in cities, *in* V. Henderson and J.-F. Thisse, eds, 'Handbook of Regional and Urban Economics', Vol. 4, North-Holland, Amsterdam, pp. 2243–2291.

Moretti, E. (2013), 'Real wage inequality', *American Economic Journal: Applied Economics* **5**(1), 65–103.

Moulton, B. R. (1990), 'An illustration of the pitfall in estimating the effects of aggregate variables on micro units', *The Review of Economics and Statistics* **72**(2), 334–338.

Stock, J. H. and Yogo, M. (2005), Testing for weak instruments in linear IV regression, *in* D. W. Andrews and J. H. Stock, eds, 'Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg', Cambridge University Press, Cambridge, pp. 80–108.

Wang, C., Wan, G. and Yang, D. (2014), 'Income inequality in the people's republic of china: Trends, determinants, and proposed remedies', *Journal of Economic Surveys* **28**(4), 686–708.

Xie, Y. and Zhou, X. (2014), 'Income inequality in today's china', *Proceedings of the National Academy of Sciences* **111**(19), 6928–6933.

Zhang, Y. and Wang, R. (2011), The main approach of proposed integrated household survey of China. presented at 4th meeting of the Wye City Group on Statistics on Rural Development and Agriculture Household Income, Brazil, 2011.

APPENDIX

A Wages and comparative statics

Low-skilled and migrant workers' wages (w_i^L and w_i^M) are given by:

$$w_i^L = s_i \left(p_{cs} A_{cs}^L\right)^\eta \left[\left(p_{cs} A_{cs}^H S_{cs}^H\right)^\rho + \left(\left(p_{cs} A_{cs}^L\right)^\eta + \left(p_{cs} A_{cs}^M S_{cs}^M\right)^\eta\right)^{\rho/\eta}\right]^{\frac{1-\rho}{\rho}} \left[\left(p_{cs} A_{cs}^L\right)^\eta + \left(p_{cs} A_{cs}^M S_{cs}^M\right)^\eta\right]^{\frac{\rho-\eta}{\eta}},$$

$$\equiv s_i \, w_{cs}^L \, .$$

$$w_i^M = s_i \left(\tfrac{A_{cs}^M}{A_{cs}^L}\right)^\eta \left(S_{cs}^M\right)^{-(1-\eta)} w_{cs}^L \equiv s_i \, w_{cs}^M \, .$$

Then, variations are given by:

$$\begin{aligned}
\frac{dw_{cs}^L}{w_{cs}^L} &= \left[\beta^L + \theta^H(1-\rho)(\beta^H - \beta^L) - \theta^M\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\beta^L-\beta^M)\right]\frac{dE_c}{E_c} \\
&+ \left[\mu^L + \theta^H(1-\rho)(\mu^H - \mu^L) - \theta^M\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\mu^L-\mu^M)\right]\frac{dS_c^H}{S_c^H} \\
&+ \left[\varphi^L + \theta^H(1-\rho)(\varphi^H - \varphi^L) - \theta^M\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\varphi^L-\varphi^M)\right]\frac{ds_c^M}{s_c^M} \\
&+ \left[\lambda^L + \theta^H(1-\rho)(1+\lambda^H - \lambda^L) - \theta^M\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\lambda^L-\lambda^M)\right]\frac{dS_{cs}^H}{S_{cs}^H} \\
&+ \left[\psi^L + \theta^H(1-\rho)(\psi^H - \psi^L) + \theta^M\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)\left(1-(\psi^L-\psi^M)\right)\right]\frac{ds_{cs}^M}{s_{cs}^M} \, .
\end{aligned}$$

$$\begin{aligned}
\frac{dw_{cs}^M}{w_{cs}^M} &= \left[\beta^M + \theta^H(1-\rho)(\beta^H - \beta^M) + \left(1-\theta^H-\theta^M\right)\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\beta^L-\beta^M)\right]\frac{dE_c}{E_c} \\
&+ \left[\mu^M + \theta^H(1-\rho)(\mu^H - \mu^M) + \left(1-\theta^H-\theta^M\right)\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\mu^L-\mu^M)\right]\frac{dS_c^H}{S_c^H} \\
&+ \left[\varphi^M + \theta^H(1-\rho)(\varphi^H - \varphi^M) + \left(1-\theta^H-\theta^M\right)\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\varphi^L-\varphi^M)\right]\frac{ds_c^M}{s_c^M} \\
&+ \left[\lambda^M + \theta^H(1-\rho)(1+\lambda^H - \lambda^M) + \left(1-\theta^H-\theta^M\right)\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)(\lambda^L-\lambda^M)\right]\frac{dS_{cs}^H}{S_{cs}^H} \\
&+ \left[\psi^M + \theta^H(1-\rho)(\psi^H - \psi^M) - \left(1-\theta^H-\theta^M\right)\left(1-\rho-\frac{\eta-\rho}{1-\theta^H}\right)\left(1-(\psi^L-\psi^M)\right)\right]\frac{ds_{cs}^M}{s_{cs}^M} \, .
\end{aligned}$$

B Data sources and definition

B.1 Main data

The mainland administrative territory of China is structured in four levels, distributed as follow for the year 2005: 31 province-level divisions (including 4 provincial-level cities or municipalities), 333 prefecture-level divisions (including 283 prefecture-level cities), 2,862 county-level divisions (including 852 districts under the jurisdiction of cities, 374 county-level cities, and 1,636 counties), and 41,636 township-level divisions (including 6,152 sub-districts or street communities, 19,522 towns, and 15,962 townships). The township-level divisions can be further divided into neighbourhood committees and village committees. A prefecture-level city typically includes a core city (also called city districts or cities proper or *shixiaqu*), as well as numerous rural counties and several county-level cities under the jurisdiction of the city government. The core city is composed of urban districts and suburban districts.

For statistical use, there are typically two types of urban/rural definitions in China, as illustrated by Zhang and Wang (2011). The first definition is based on the household registration status. Before the reforms started in 1978, residents in neighbourhood committees were given a non-agricultural *Hukou*, while those in village committees were given an agricultural *Hukou*. Therefore, all neighbourhood committees were defined as urban areas and all village committees as rural areas. The second definition is closely related to China's actual urbanisation process and was introduced in the 2000 China Population Census. As highlighted by Chan and Hu (2003), three new elements have been introduced to define urban areas (city and town areas) at various geographic levels: (a) whether the area has an average population density of 1,500 persons/sq.km; (b) whether the local government is located in the area; and (c) whether the area is connected to an area where the local government is located.

For our analysis, we consider only the urban area of the core- and county-level cities among the 284 provincial- or prefecture-level cities.[15] The city area includes all township-level units in the core city with an average population density of at least 1,500 persons per sq.km. For core cities with a density below 1,500 persons per sq.km or county-level cities, the city area includes: (a) the township-level unit (street, town and township) where the district or city government is located; (b) township-level units with a built-up area contiguous to (a); and (c) all other streets. This is largely comparable to a metropolitan area in Western country standards.

The micro data we use are drawn from a random extraction of 20% of the *1% 2005 China Population Survey* conducted by the National Bureau of Statistics of China. The survey adopts a stratified multi-stage PPS cluster sampling method and provides information on all respondents about their socio-economic characteristics, family structure, employment status, wages, housing attributes, social insurance, etc. A major feature of the survey is that it includes not only registered urban residents but also migrants. Migrants are defined as individuals who have been living in a

[15] Urumqi, Karamay, and Lhasa are excluded.

county-level region different from their *Hukou* registration county-level region (rural counties, urban districts inside the core cities, or county-level cities) for more than 6 months (conditional on the 2 counties not being in the same core city). Migrants from both urban and rural areas are included. Other important features of the survey are the availability of (a) hourly labour earnings for wage workers at the individual level; (b) a two-digit industrial classification (GB/T4754-2002) that allows us to disentangle 36 industries as detailed in Combes et al. (2019), Table 1; and (c) a large number of observations that allows us to run estimations for three groups of workers separately. The sample available to us includes 2,585,481 individuals distributed throughout China.

The sample is first restricted to 1,353,529 individuals aged 16–70 who report (positive) earnings. We then retain those who are non-agricultural workers and live in the cities defined above, which leaves us with 315,779 individuals. The last restriction we apply is that for any industry, the three worker categories defined in the text (high-skilled natives, low-skilled natives and rural migrants) are present in all cities such that worker-type-specific estimates refer to the same group of city-industries. The final sample is then composed of 245,935 workers located in 257 cities and 36 industries, for a total of 2,554 city-industries.

B.2 Migrant instruments

The instrument used in Appendix F Table 8 is based on M_c^{1990}, the actual number of migrants in city c in 1990, which is divided by the number of low-skilled workers in city c in 1990, L_c^{1990} :

$$\frac{M_c^{1990}}{L_c^{1990}}. \tag{12}$$

Let M_{pc}^{9500} be the number of migrants from province p to city c over 1995–2000. For the instrument used in Table 2, we first estimate a gravity model such that

$$M_{pc}^{9500} = a_p + b_c + g.dist_{pc} + d_{pc} + f_{pc} + e_{pc}, \tag{13}$$

where a_p and b_c are province and city fixed effects, $dist_{pc}$ is the distance as the crow flies between p and c and g its effect, d_{pc} is the effect of a dummy variable equal to 1 when c is in province p, and f_{pc} is the effect of a dummy variable when c is in a province contiguous to p and e_{pc} is a random component. This estimation is based on 1,610 observations and yields an R^2 of 0.64. The distance effect is -0.31, and the two dummy effects are 1.63 and 0.5195, respectively. The three effects are significantly different from zero at the 1% level.

This allows us to obtain a predicted migrant flow as follows:

$$\widehat{M}_{pc}^{9500} = \hat{a}_p + \hat{b}_c + \hat{g}\, dist_{pc} + \hat{d}_{pc} + \hat{f}_{pc}. \tag{14}$$

Then, one can compute the share of migrants to city c in province p predicted from the gravity

model by dividing the predicted flow of migrants by its sum over all destination cities:

$$\widehat{ShM}_{pc}^{9500} = \frac{\widehat{M}_{pc}^{9500}}{\sum_c \widehat{M}_{pc}^{9500}}. \tag{15}$$

Compared to the actual share of migrants we have in the data, this removes the part that does not relates to proximity and origin and destination fixed features, which is the part that is possibly the most endogeneous. Then we also reduce the possible endogeneity of the instrument by not multiplying this share by the actual number of migrants from p but simply by province p total employment, again removing some of the idiosyncrasies that could make migrant flows from a particular province larger or lower compared to its size. We then obtain a prediction for the total number of migrants in c by summing over all provinces:

$$\widehat{M}_c^{9500} = \sum_p \widehat{ShM}_{pc}^{9500} E_p^{00}, \tag{16}$$

where E_p^{00} is the province employment in 2000. The instrument used in Table 2 is

$$\frac{\widehat{M}_c^{9500}}{E_c^{00}}, \tag{17}$$

where E_c^{00} is city c employment in 2000.

Then the same instruments can be computed at the industry level by first running the gravity equation on each industry separately and then by finally diving by employment in the city-industry in 2000.

Table 4 – Definition of variables

Variable name	Formula	Definition
Individual variables		
Male		Dummy variable: Male = 1
Years of education		Number of years of schooling (minus 12 for high-skilled natives, minus 6 for low-skilled natives and migrants)
Experience	Age − Education − 6	Potential experience (in years)
Head of institution		Dummy variable: Head of the institution= 1
Technical worker		Dummy variable: Technical workers= 1
Office worker		Dummy variable: Office worker= 1
Service worker		Dummy variable: Service workers= 1
Self-employed		Dummy variable: Self-employed workers= 1
Private firm		Dummy variable: Private and individual enterprises = 1
City-industry variables		
Migrant/Low$_{cs}$ (S^M_{cs})	$\frac{\mathrm{M}_{cs}}{\mathrm{L}_{cs}}$	Ratio of migrants to unskilled workers at the city-industry level
High/Low$_{cs}$ (S^H_{cs})	$\frac{\mathrm{H}_{cs}}{\mathrm{L}_{cs}}$	Ratio of skilled to unskilled workers at the city-industry level
Specialisation$_{cs}$	$\frac{\mathrm{E}_{cs}}{\mathrm{E}_{c}}$	Share of the industry s in local total employment
City level variables		
Migrant/Low$_{c}$ (S^M_{c})	$\frac{\mathrm{M}_{c}}{\mathrm{L}_{c}}$	Ratio of migrants to unskilled workers at the city level
Density$_c$	$\frac{\mathrm{E}_{c}}{\mathrm{Area}_{c}}$	Total employment density in the city
Land area$_c$		Land area of the city
High/Low$_{c}$ (S^H_{c})	$\frac{\mathrm{H}_{c}}{\mathrm{L}_{c}}$	Ratio of skilled to unskilled workers at the city level
Market access$_c$	$\sum_{\text{city } c' \neq c} \frac{\mathrm{E}_{c'}}{\mathrm{Distance}_{cc'}}$	Sum of employment in other Chinese cities c' inversely weighted by the distance as the crow flies to the city c considered
Port proximity$_c$		Proximity to the closest seaport in terms of the volume of freight handled, (in decreasing order: Shanghai, Ningbo, Guangzhou, Tianjin, Qingdao, Qinhuangdao, Dalian, Rizhao, Yingkou, Yantai, Lianyungang, Zhanjiang).
Diversity$_c$	$\left[\sum_s \left(\frac{\mathrm{E}_{cs}}{\mathrm{E}_{c}}\right)^2\right]^{-1}$	Inverse-Herfindhal index on employment
City-industry instruments		
1990 Migrant/Low$_{cs}$		1990 Ratio of migrants to unskilled workers at city-industry level
2000 Pred. migrant/ Emp.$_{cs}$		2000 Migrant ratio over employment, predicted from a province-city gravity model at city-industry level
1990 High/Low$_{cs}$		1990 Ratio of skilled to unskilled workers at city-industry level
1990 Specialisation$_{cs}$		1990 Share of the industry s in local total employment
1995 SOE share$_{cs}$		1995 State-owned enterprises share at city-industry level
City instruments		
1964 Density$_c$		1964 Population density
1964 Non-ag. share in emp$_c$		1964 Non-agricultural population share
1982 Land area$_c$		1982 Land size
1982 Ag. share in emp$_c$		1982 Agricultural population share
1982 Nb of college students$_c$		1982 Number of college students per 10,000 persons
2000 Pred. migrant/ Emp.$_c$		2000 Migrant ratio over employment, predicted from a province-city gravity model at the city level (see Appendix A.2)
1990 Migrant/Low$_c$		1990 Ratio of migrants to unskilled workers at the city level

Notes: In the regressions, the logarithm of the variable is used for all variables. For city-industry variables, their logarithm is then centered with respect to the city mean for each corresponding sample (high-skilled natives, low-skilled natives and rural migrants).

C Individual wages and location

Table 5 – Individual characteristics marginal effects (Step 1)

	High-skilled urban natives	Low-skilled urban natives	Rural migrants
Male	0.0925^a	0.204^a	0.130^a
	(0.00536)	(0.00384)	(0.00473)
Years of education	0.199^a	0.0517^a	0.0566^a
	(0.00365)	(0.00105)	(0.00123)
Potential experience	0.0260^a	0.0146^a	0.0123^a
	(0.000989)	(0.000583)	(0.000667)
Potential experience squared	−0.000440^a	−0.000317^a	-0.000271^a
	(0.0000288)	(0.0000128)	(0.0000172)
Managerial work	0.356^a	0.500^a	0.622^a
	(0.0133)	((0.0123)	((0.0235)
Technical work	0.158^a	0.215^a	0.293^a
	(0.0104)	(0.00751)	((0.0183)
Office work	0.173^a	0.0618^a	0.127^a
	(0.0108)	(0.00726)	((0.0125)
Service work	0.0232	-0.0161^a	0.0487^a
	(0.0124)	(0.00551)	(0.00875)
Self-employment	0.0597^a	0.156^a	0.0865^a
	(0.0174)	(0.00563)	(0.00825)
Private sector	0.0322^a	−0.151^a	−0.0777^a
	(0.00794)	(0.00453)	(0.00845)
Observations	50,834	127,750	67,351
R^2	0.16	0.14	0.14

Notes: OLS within city-industry estimator for equation (9). City-industry fixed effects for the second step are backed up using Frish-Waugh theorem. See Appendix B, Table 4 for a definition of variables. Standard errors in parentheses. ^{b}: p<0.05, ^{a}: p<0.01.

D List of instrumental variables for Steps 2 and 3

Table 6 – List of instrumental variables

	High-Skilled urban natives		Low-Skilled urban natives		Rural migrants	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a) - Table 2 (step 2)		IV		IV		IV
1990 Migrant/Low-skilled$_{cs}$		Y		N		Y
2000 (Predicted migrant)/ Employment$_{cs}$		N		Y		Y
1990 High/Low-skilled$_{cs}$		Y		Y		Y
1990 Specialisation$_{cs}$		Y		Y		Y
1995 SOE share$_{cs}$		Y		Y		Y
Panel (b) - Table 2 (step 3)	IV1	IV2	IV1	IV2	IV1	IV2
1964 Employment density$_{c}$	N	Y	Y	Y	N	N
1964 Non-agricultural share in employment$_{c}$	Y	N	N	N	Y	Y
1982 Land area$_{c}$	Y	Y	Y	Y	Y	Y
1982 Agricultural share in employment$_{c}$	N	Y	Y	Y	N	N
1982 College students per 10,000 persons$_{c}$	Y	N	N	N	Y	Y
2000 (Predicted migrant)/ Employment$_{c}$	Y	Y	Y	Y	Y	Y
Panel (c) - Appendix Table 8 (step 3)	IV1	IV2	IV1	IV2	IV1	IV2
1964 Density$_{c}$	N	Y	Y	Y	Y	Y
1964 Non-agricultural share$_{c}$	Y	N	N	N	N	N
1982 Land area$_{c}$	Y	Y	Y	Y	Y	Y
1982 Agricultural share$_{c}$	N	Y	Y	Y	Y	Y
1982 College students per 10,000 persons$_{c}$	Y	N	N	N	N	N
1990 Migrant/ Low-skilled$_{c}$	Y	Y	Y	Y	Y	Y

Notes: See Appendix B, Table 4 for a definition of variables. Panel (a): City-industry characteristics impact. Panel (b): City characteristics impact.

E Magnitude of urban effects

Table 7 – Magnitude of urban effects

	High-Skilled urban natives		Low-Skilled urban natives		Rural migrants	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a)	P75-P25	%	P75-P25	%	P75-P25	%
Migrant/Low-skilled$_{cs}$	1.44	0.0	1.13	−2.0	0.87	−2.2
High/Low-skilled$_{cs}$	1.73	7.2	1.00	3.5	0.91	2.1
Specialisation$_{cs}$	0.89	1.0	1.05	3.8	1.07	1.4
Panel (b)	P75-P25	%	P75-P25	%	P75-P25	%
Migrant/Low$_c$	1.26	27.7	1.74	23.8	1.40	7.7
Density$_c$	1.02	6.9	1.16	7.8	1.09	3.1
Land area$_c$	1.39	16.4	1.14	8.3	0.85	1.6
High/Low$_c$	0.55	3.5	0.73	1.7	0.93	3.8
Market access$_c$	0.35	7.0	0.37	8.2	0.29	3.1
Port proximity$_c$	1.85	4.1	1.90	3.6	1.44	4.0
Diversity$_c$	0.20	0.0	0.30	−4.2	0.44	4.5

Notes: See Appendix B, Table 4 for a definition of variables. Panel (a): City-industry characteristics impact. Panel (b): City characteristics impact. Columns 'P75-P25' report the difference between the 75th and 25th percentiles of the variable used in the estimation, that is, the logarithm of the city-industry characteristic centered with respect to its city mean for the corresponding sample in the second step and the logarithm of the city variable in the first step. Then this difference is multiplied by the OLS point estimate of the effect and the exponential is finally taken to obtain the impact of the variable.

F Supplementary instrumented regressions for Step 3

Table 8 – Supplementary IV for Step 3

	High-skilled urban natives		Low-skilled urban natives		Rural migrants	
	(1)	(2)	(3)	(4)	(5)	(6)
	IV1	IV2	IV1	IV2	IV1	IV2
Migrant/Low$_c$	0.184^a	0.156^a	0.0934^a	0.0782^a	-0.00580	-0.0166
	(0.0271)	(0.0359)	(0.0223)	(0.0233)	(0.0167)	(0.0228)
Density$_c$	0.181^a	0.144^a	0.155^a	0.0728^b	0.132^a	0.0655^b
	(0.0388)	(0.0541)	(0.0270)	(0.0299)	(0.0268)	(0.0289)
Land area$_c$	0.0937^a	0.0957^a	0.0731^a	0.0812^a	0.138^a	0.0896^a
	(0.0214)	(0.0352)	(0.0192)	(0.0249)	(0.0146)	(0.0311)
High/Low$_c$		0.0835^b		0.0566^b		0.0693^a
		(0.0331)		(0.0223)		(0.0221)
Market access$_c$		0.125^b		0.260^a		0.219^a
		(0.0488)		(0.0467)		(0.0418)
Port proximity$_c$		0.0265^a		0.0238^a		0.0253^a
		(0.00710)		(0.00577)		(0.00625)
Diversity$_c$		-0.0836		-0.185^a		-0.127
		(0.0944)		(0.0587)		(0.0695)
Observations	257	257	257	257	257	257
Weakness	15.5	8.3	17.5	9.9	15.6	13.3
Overidentification	0.07	0.19	0.25	0.32	0.45	0.29

Notes: Estimations performed with LIML. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. See Appendix B, Table 4 for a definition of variables. Density, land area and the rural to low-skilled workers ratio are instrumented. Other variables are introduced as controls in columns 'IV2' only. Instruments at city-industry level and city-level are defined in Appendix B and those used for each column listed in Table 6, Appendix D. Line 'overidentification' reports the p-value for the Hansen J-test; Line 'weakness' reports the Cragg-Donald statistics. Following Stock and Yogo (2005), critical values for LIML estimates at the 5% level are below 5 for any of the estimations reported in the table.

Unequal Migration and Urbanisation Gains in China

Supplementary Material [1]

Pierre-Philippe Combes [2] Sylvie Démurger [3] Shi Li [4] Jianguo Wang [5]

January 17, 2019

Contents

[1]Univ Lyon, CNRS, GATE UMR 5824, F-69130 Ecully, France; Sciences Po, Department of Economics, 28, Rue des Saints-Pères, 75007 Paris, France; Also research fellow at the CEPR. Email: ppcombes@gmail.com.

[2]Corresponding author. Univ Lyon, CNRS, GATE UMR 5824, F-69130 Ecully, France. Also research fellow at IZA. Email: sylvie.demurger@cnrs.fr.

[3]Business School, Beijing Normal University, China. Also research fellow at IZA, Bonn, Germany. Email: lishi@bnu.edu.cn.

[4]Beijing Information Science and Technology University, China. Email: jgwang0225@gmail.com.

1 Definition of industries

Table 1 – Definition of the 36 industries

No.	Industry name	Sub-categories
1	Mining	Coal mining and dressing (06); Petroleum and natural gas extraction (07); Ferrous metal ores mining (08) ;Non-ferrous metal ores mining (09); Non-metal ores mining (10); Other ores mining (11)
2	Utilities	Electric and heat power supply (44); Gas supply (45); Tap water supply (46)
3	Food, beverages, tobacco	Agricultural products processing (13); Food processing (14); Alcohol, beverages, and refined tea production(15); Tobacco processing (16)
4	Textiles	Textiles production (17)
5	Garment and leather	Garment, shoe, and head-wear tailoring (18); Leather, fur, feather and related products (19)
6	Wood, paper and furniture	Timber processing, and wood, bamboo, rattan, palm, and straw products (20); Furniture industries (21); Paper making and paper products (22)
7	Printing and culture	Printing and recorded media (23); Culture, education, art, sports, and entertainment products making (24); Art craft and other products (42)
8	Petroleum and chemicals	Petroleum processing, coking, nuclear fuel processing (25); Chemical raw materials and chemical products (26); Manufacture of medicines (27); Chemical fibres production (28); Waste resources recycling (43)
9	Plastic and rubber	Rubber products (29); Plastics products (30)
10	Clay and glass	Non-metallic mineral products (31)
11	Metal products	Ferrous metal smelting and processing (32); Non-ferrous metal smelting and processing (33); Metal products (34)
12	General machinery	General purpose machinery production (35)
13	Special machinery	Special purpose machinery production (36)
14	Transportation equipment	Transportation equipment production (37)
15	Electric equipment	Electrical machinery and equipment production (39)
16	Electronic equipment	Communication, computers and other electronic equipment production (40); Measuring instruments production (41)
17	Construction	Housing and Civil engineering (47)
18	Renovation and decoration	Renovation (48); Decoration (49); Other construction (50)
19	Railway transport	Railway transport (51)
20	Road transport	Road transport (52)
21	Other transport	City public transport (53); Water transport (54); Air transport (55); Pipeline transport (56); Loading, unloading, removal, and other transport services (57); Storage (58);Postal services (59)
22	Telecom and IT	Telecommunications, radio, television, and satellite transmission services (60); Computer services (61); software services (62)
23	Wholesale trade	Wholesale trade (63)
24	Retail trade	Retail trade (65)
25	Hotel	Accommodation (66)
26	Catering	Catering (67)
27	Finance and insurance	Banking (68); Securities (69); Insurance (70); Other financial activities (71)
28	Real estate	Real estate (72)
29	Leasing	Leasing (73); Commercial services (74)
30	Research and polytechnic services	Research and experimental development (75); Polytechnic services (76); Science and technology promotion and application (77); Geological exploration (78)
31	Water, environment, public facilities	Water management (79); Ecological protection and environmental management (80); Management of public facilities (81)
32	Resident services	Resident services (82); Other services (83)
33	Education	Education (84)
34	Health, social insurance and welfare	Health care (85); Social Insurance (86); Social welfare (87)
35	Culture, sports and entertainment	News and publishing (88); Radio, television, film, and television recording and production (89); Culture and arts (90); Sports (91); Entertainment (92)
36	Administration, social organizations	Chinese Communist Party organs (93); State institutions (94); People's Political Consultative Conference and democratic parties (95); Mass and social organizations, and other membership organizations (96); Autonomous grass-roots organisations (97); International organizations (98)

2 Descriptive statistics

Table 2 – Summary statistics for city characteristics, High-skilled natives

	Mean	Std. Dev.	p10	p50	p90
Density (workers per sq. km)	824.1	703.3	265.3	529.6	1500.1
Land area (sq. km)	4192.5	3515.7	617	3257	12188
Migrant/Low	0.63	0.96	0.095	0.33	1.02
High/Low	0.54	0.23	0.27	0.50	0.96
Market access	207114.6	56290.1	127502.3	210686	272265.1
Port proximity	0.070	0.12	0.0011	0.0051	0.30
Diversification	18.4	3.35	13.0	19.0	22.8

Notes: 257 observations. Statistics computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 3 – Summary statistics for city-industry characteristics, High-skilled natives

	Mean	Std. Dev.	p10	p50	p90
Migrant/Low	0.49	1.19	0.037	0.18	0.95
High/Low	1.40	1.49	0.16	0.89	3.15
Specialisation	0.054	0.046	0.015	0.039	0.13

Notes: 2,554 observations. Statistics computed on the weighted number of observations in the city-industry for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 4 – Summary statistics for city characteristics, Low-skilled natives

	Mean	Std. Dev.	p10	p50	p90
Density (workers per sq. km)	754.8	628.2	186.7	575.3	1500.1
Land area (sq. km)	3212.0	2900.3	530	2258	7152
Migrant/Low	0.46	0.72	0.059	0.24	1.02
High/Low	0.42	0.21	0.18	0.40	0.63
Market access	209585.9	60000.9	127502.3	211376.0	275997.6
Port proximity	0.061	0.12	0.0011	0.0048	0.30
Diversification	17.4	3.70	12.2	18.1	21.7

Notes: 257 observations. Statistics computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 5 – Summary statistics for city-industry characteristics, Low-skilled natives

	Mean	Std. Dev.	p10	p50	p90
Migrant/Low	0.49	1.05	0.040	0.20	1.02
High/Low	0.37	0.62	0.038	0.17	0.86
Specialisation	0.073	0.061	0.017	0.049	0.16

Notes: 2,554 observations. Statistics computed on the weighted number of observations in the city-industry for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 6 – Summary statistics for city characteristics, Rural migrants

	Mean	Std. Dev.	p10	p50	p90
Density (workers per sq. km)	1269.5	1061.9	309.4	848.7	3299.1
Land area (sq. km)	3319.4	2794.7	968	2176	6543
Migrant/Low	1.55	1.79	0.19	0.77	4.97
High/Low	0.51	0.25	0.20	0.49	0.83
Market access	231824.4	60224.2	156032.8	226611.3	285602.8
Port proximity	0.075	0.12	0.0016	0.010	0.30
Diversification	17.1	3.92	13.0	17.6	22.8

Notes: 257 observations. Statistics computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 7 – Summary statistics for city-industry characteristics, Rural migrants

	Mean	Std. Dev.	p10	p50	p90
Migrant/Low	2.75	3.93	0.22	1.14	9.95
High/Low	0.37	0.50	0.058	0.24	0.88
Specialisation	0.074	0.060	0.017	0.051	0.16

Notes: 2,554 observations. Statistics computed on the weighted number of observations in the city-industry for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

3 Correlations

Table 8 – Simple correlations between city characteristics, High-skilled natives

	(2)	(3)	(4)	(5)	(6)	(7)
Migrant/Low-skilled (1)	0.48	0.28	0.42	0.31	0.36	0.12
Density (2)	1	-0.36	0.14	0.43	0.28	0.17
Land area (3)		1	0.40	-0.09	0.30	0.45
High/Low-skilled (4)			1	-0.16	-0.01	0.29
Market Access (5)				1	0.47	0.22
Port proximity (6)					1	0.42
Diversity (7)						1

Notes: Correlations computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 9 – Simple correlations between migrant ratio, density and land area with city instruments, High-skilled natives

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Migrant/Low (1)	0.48	0.28	0.13	-0.07	-0.22	-0.25	0.28	0.64	0.76
Density (2)	1	-0.36	0.28	0.01	-0.50	-0.25	0.20	0.34	0.25
Land area (3)		1	0.34	0.32	0.41	-0.41	0.48	0.06	0.22
1964 Density (4)			1	0.52	-0.21	-0.76	0.49	-0.27	-0.06
1964 Non-agricultural share (5)				1	0.20	-0.49	0.37	-0.24	-0.07
1982 Land area (6)					1	0.28	0.06	-0.22	-0.06
1982 Agricultural share (7)						1	-0.52	0.08	-0.07
1982 College students (8)							1	0.04	0.10
1990 Migrants/Low (9)								1	0.56
2000 Predicted migrants/Employment (10)									1

Notes: Correlations computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 10 – Simple correlations between city-industry characteristics and instruments, High-skilled natives

	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Migrant/Low (1)	0.07	-0.36	0.33	0.00	-0.39	0.12	-0.09
Specialisation (2)	1	-0.11	0.16	0.54	-0.05	-0.02	-0.08
High/Low (3)		1	-0.31	-0.13	0.68	0.08	0.12
1990 Migrant/Low (4)			1	0.12	-0.33	0.04	-0.05
1990 Specialisation (5)				1	-0.10	-0.07	-0.08
1990 High/Low (6)					1	0.10	0.09
2000 Predicted migrant/employment (7)						1	-0.01
1995 SOE share (8)							1

Notes: Correlations computed on the weighted number of observations in the city-industry for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 11 – Simple correlations between city characteristics, Low-skilled natives

	(2)	(3)	(4)	(5)	(6)	(7)
Migrant/Low-skilled (1)	0.42	0.31	0.36	0.32	0.38	0.26
Density (2)	1	-0.38	0.06	0.46	0.30	0.21
Land area (3)		1	0.37	-0.13	0.25	0.43
High/Low-skilled (4)			1	-0.16	0.07	0.47
Market Access (5)				1	0.43	0.21
Port proximity (6)					1	0.42
Diversity (7)						1

Notes: Correlations computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 12 – Simple correlations between migrant ratio, density and land area with city instruments, Low-skilled natives

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Migrant/Low (1)	0.42	0.31	0.25	0.12	-0.13	-0.31	0.32	0.56	0.71
Density (2)	1	-0.38	0.45	0.15	-0.42	-0.28	0.24	0.19	0.28
Land area (3)		1	0.26	0.33	0.34	-0.37	0.40	0.16	0.15
1964 Density (4)			1	0.46	-0.32	-0.66	0.49	-0.11	0.05
1964 Non-agricultural share (5)				1	0.03	-0.57	0.50	0.07	0.05
1982 Land area (6)					1	0.29	-0.11	-0.07	-0.10
1982 Agricultural share (7)						1	-0.62	-0.09	-0.12
1982 College students (8)							1	0.20	0.08
1990 Migrants/Low (9)								1	0.49
2000 Predicted migrants/Employment (10)									1

Notes: Correlations computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 13 – Simple correlations between city-industry characteristics and instruments, Low-skilled natives

	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Migrant/Low (1)	0.16	-0.23	0.30	0.01	-0.30	0.24	-0.10
Specialisation (2)	1	-0.20	0.12	0.58	-0.11	-0.09	-0.15
High/Low (3)		1	-0.20	-0.12	0.61	-0.01	0.18
1990 Migrant/Low (4)			1	-0.01	-0.23	0.11	-0.09
1990 Specialisation (5)				1	-0.19	-0.18	-0.08
1990 High/Low (6)					1	-0.02	0.18
2000 Predicted migrant/employment (7)						1	-0.02
1995 SOE share (8)							1

Notes: Correlations computed on the weighted number of observations in the city-industry for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 14 – Simple correlations between city characteristics, Rural migrants

	(2)	(3)	(4)	(5)	(6)	(7)
Migrant/Low-skilled (1)	0.63	0.03	0.38	0.34	0.17	-0.36
Density (2)	1	-0.32	0.31	0.27	0.18	-0.22
Land area (3)		1	0.34	-0.05	0.39	0.53
High/Low-skilled (4)			1	-0.21	0.04	0.15
Market Access (5)				1	0.42	0.09
Port proximity (6)					1	0.45
Diversity (7)						1

Notes: Correlations computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 15 – Simple correlations between migrant ratio, density and land area with city instruments, Rural migrants

	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Migrant/Low (1)	0.63	0.03	-0.32	-0.52	-0.52	0.12	-0.04	0.82	0.86
Density (2)	1	-0.32	-0.08	-0.31	-0.64	-0.08	0.20	0.54	0.44
Land area (3)		1	0.35	0.36	0.25	-0.47	0.43	-0.02	0.09
1964 Density (4)			1	0.71	0.07	-0.75	0.40	-0.57	-0.38
1964 Non-agricultural share (5)				1	0.33	-0.57	0.42	-0.62	-0.47
1982 Land area (6)					1	0.18	-0.05	-0.53	-0.39
1982 Agricultural share (7)						1	-0.59	0.25	0.22
1982 College students (8)							1	-0.08	-0.13
1990 Migrants/Low (9)								1	0.75
2000 Predicted migrants/Employment (10)									1

Notes: Correlations computed on the weighted number of observations in the city for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

Table 16 – Simple correlations between city-industry characteristics and instruments, Rural migrants

	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Migrant/Low (1)	0.21	-0.06	0.27	-0.07	-0.21	0.33	-0.09
Specialisation (2)	1	-0.16	0.04	0.45	0.04	-0.06	-0.19
High/Low (3)		1	-0.14	0.04	0.46	-0.04	0.25
1990 Migrant/Low (4)			1	0.04	-0.23	0.13	-0.14
1990 Specialisation (5)				1	-0.01	-0.22	-0.14
1990 High/Low (6)					1	-0.10	0.33
2000 Predicted migrant/employment (7)						1	-0.03
1995 SOE share (8)							1

Notes: Correlations computed on the weighted number of observations in the city-industry for the corresponding group of workers. See Appendix B, Table 4 for a definition of variables.

4 First-stage regressions for IV estimates in Table 2

Table 17 – First-stage regressions for IV estimates - Step 2

	High-skilled natives			Low-skilled natives			Rural migrants		
	Mig.	High	Specia.	Mig.	High	Specia.	Mig.	High	Specia.
1990 Specia.$_{cs}$	0.00570	-0.0130	0.221^{a}	-0.0146	-0.0283	0.278^{a}	-0.0565^{a}	-0.0101	0.290^{a}
	(0.0228)	(0.0141)	(0.0145)	(0.0227)	(0.0170)	(0.0153)	(0.0194)	(0.0171)	(0.0185)
1990 High/Low-skilled$_{cs}$	-0.0796^{a}	0.0800^{a}	-0.00925	-0.0718^{a}	0.102^{a}	0.0501^{a}	-0.0378^{b}	0.0346^{b}	0.0454^{b}
	(0.0251)	(0.0155)	(0.0159)	(0.0242)	(0.0181)	(0.0163)	(0.0185)	(0.0163)	(0.0176)
1990 Mig./Low-skilled$_{cs}$	0.0770^{a}	0.00744	0.0749^{a}				0.124^{a}	0.0533^{a}	0.0462^{a}
	(0.0167)	(0.0103)	(0.0106)				(0.0119)	(0.0105)	(0.0114)
1995 SOE share$_{cs}$	0.245^{b}	0.126^{b}	-0.0632	0.250^{a}	0.0777	-0.0374	0.0830^{b}	0.107^{a}	-0.168^{a}
	(0.0959)	(0.0593)	(0.0608)	(0.0745)	(0.0556)	(0.0501)	(0.0419)	(0.0369)	(0.0400)
2000 Pred. mig./ Emp$_{cs}$				0.152^{a}	0.00866	0.000257	0.167^{a}	-0.0195	0.0288
				(0.0179)	(0.0134)	(0.0121)	(0.0196)	(0.0172)	(0.0187)
N	2,554	2,554	2,554	2,554	2,554	2,554	2,554	2,554	2,554
R^2	0.51	0.84	0.62	0.38	0.67	0.63	0.51	0.65	0.61

Notes: See Table 2. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B.

Table 18 – First-stage regressions for IV estimates - Step 3, High-skilled natives

	(1) IV1 Mig.	(2) IV1 Dens.	(3) IV1 Area	(4) IV2 Mig.	(5) IV2 Dens.	(6) IV2 Area
1964 Employment density$_c$				0.0204	0.0933	0.0917
				(0.0524)	(0.0554)	(0.0594)
1964 Non-agricultural share in employment$_c$	-0.124^b	0.0733	0.165^b			
	(0.0624)	(0.0681)	(0.0766)			
1982 Land area$_c$	-0.152^a	-0.359^a	0.315^a	-0.153^a	-0.355^a	0.416^a
	(0.0334)	(0.0364)	(0.0410)	(0.0417)	(0.0442)	(0.0473)
1982 Agricultural share in employment$_c$				-0.0165	0.179^b	-0.354^a
				(0.0664)	(0.0702)	(0.0753)
1982 College students per 10,000 persons$_c$	0.376^a	0.223^a	0.566^a			
	(0.0611)	(0.0666)	(0.0749)			
2000 (Predicted migrant)/ Employment$_c$	1.122^a	0.253^a	0.315^a	0.955^a	0.104	0.298^a
	(0.0586)	(0.0639)	(0.0719)	(0.0725)	(0.0767)	(0.0822)
High/Low$_c$				0.543^a	0.440^a	0.160
				(0.101)	(0.107)	(0.115)
Market access$_c$				0.287	0.452^b	-0.277
				(0.174)	(0.184)	(0.197)
Port proximity$_c$				0.0486	0.0398	0.0332
				(0.0268)	(0.0284)	(0.0304)
Diversity$_c$				-0.0990	0.400	0.343
				(0.271)	(0.287)	(0.307)
N	257	257	257	257	257	257
R^2	0.66	0.35	0.42	0.67	0.41	0.53

Notes: See Table 2. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B.

Table 19 – First-stage regressions for IV estimates - Step 3, Low-skilled natives

	(1) IV1 Mig.	(2) IV1 Dens.	(3) IV1 Area	(4) IV2 Mig.	(5) IV2 Dens.	(6) IV2 Area
1964 Employment density$_c$	0.107^b	0.287^a	0.129^b	0.0911	0.249^a	0.105
	(0.0489)	(0.0477)	(0.0522)	(0.0536)	(0.0534)	(0.0580)
1982 Land area$_c$	0.0169	-0.217^a	0.436^a	-0.00285	-0.190^a	0.343^a
	(0.0414)	(0.0403)	(0.0442)	(0.0470)	(0.0469)	(0.0509)
1982 Agricultural share in employment$_c$	-0.174^a	0.0966	-0.407^a	-0.0999	0.101	-0.288^a
	(0.0647)	(0.0631)	(0.0691)	(0.0715)	(0.0713)	(0.0774)
2000 (Predicted migrant)/ Employment$_c$	1.138^a	0.313^a	0.213^a	0.997^a	0.219^a	0.207^b
	(0.0693)	(0.0676)	(0.0740)	(0.0773)	(0.0771)	(0.0837)
High/Low$_c$				0.436^a	0.179	0.0211
				(0.0989)	(0.0985)	(0.107)
Market access$_c$				0.477^b	0.571^a	-0.525^b
				(0.190)	(0.189)	(0.206)
Port proximity$_c$				0.0186	-0.000281	0.0348
				(0.0303)	(0.0302)	(0.0328)
Diversity$_c$				-0.360	-0.147	0.745^a
				(0.256)	(0.255)	(0.277)
N	257	257	257	257	257	257
R^2	0.57	0.35	0.38	0.60	0.37	0.42

Notes: See Table 2. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B.

Table 20 – First-stage regressions for IV estimates - Step 3, Rural migrants

	(1) IV1 Mig.	(2) IV1 Dens.	(3) IV1 Area	(4) IV2 Mig.	(5) IV2 Dens.	(6) IV2 Area
1964 Non-agricultural share in employment$_c$	-0.255^a	-0.190^a	0.270^a	-0.0859	-0.0704	0.0193
	(0.0529)	(0.0636)	(0.0645)	(0.0626)	(0.0783)	(0.0770)
1982 Land area$_c$	-0.166^a	-0.350^a	0.212^a	-0.142^a	-0.317^a	0.189^a
	(0.0297)	(0.0357)	(0.0362)	(0.0287)	(0.0360)	(0.0353)
1982 College students per 10,000 persons$_c$	0.203^a	0.345^a	0.433^a	0.100	0.364^a	0.262^a
	(0.0522)	(0.0627)	(0.0636)	(0.0730)	(0.0914)	(0.0898)
2000 (Predicted migrant)/ Employment$_c$	1.342^a	0.283^a	0.496^a	1.085^a	0.0651	0.428^a
	(0.0610)	(0.0732)	(0.0744)	(0.0651)	(0.0815)	(0.0801)
High/Low$_c$				0.438^a	0.249^b	0.0490
				(0.0905)	(0.113)	(0.111)
Market access$_c$				0.681^a	0.412^b	-0.389^b
				(0.140)	(0.175)	(0.172)
Port proximity$_c$				0.0636^a	0.104^a	0.0973^a
				(0.0211)	(0.0264)	(0.0259)
Diversity$_c$				-1.096^a	-1.281^a	1.137^a
				(0.191)	(0.239)	(0.235)
N	257	257	257	257	257	257
R^2	0.80	0.51	0.38	0.84	0.58	0.50

Notes: See Table 2. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B.

5 First-stage regressions for IV estimates in Table 8 - Step 3

Table 21 – First-stage regressions for IV estimates - Step 3, High-skilled natives

	(1) IV1 Mig.	(2) IV1 Dens.	(3) IV1 Area	(4) IV2 Mig.	(5) IV2 Dens.	(6) IV2 Area
1964 Employment density$_c$				0.115	0.156^a	0.165^a
				(0.0614)	(0.0579)	(0.0620)
1964 Non-agricultural share in employment$_c$	-0.00911	0.124	0.191^b			
	(0.0780)	(0.0689)	(0.0798)			
1982 Land area$_c$	-0.0872^b	-0.335^a	0.330^a	-0.0378	-0.311^a	0.479^a
	(0.0413)	(0.0365)	(0.0423)	(0.0482)	(0.0455)	(0.0488)
1982 Agricultural share in employment$_c$				-0.00453	0.173^b	-0.356^a
				(0.0730)	(0.0689)	(0.0738)
1982 College students per 10,000 persons$_c$	0.397^a	0.214^a	0.576^a			
	(0.0745)	(0.0658)	(0.0763)			
1990 Migrant/ Low-skilled$_c$	0.617^a	0.203^a	0.158^a	0.536^a	0.172^a	0.263^a
	(0.0481)	(0.0425)	(0.0493)	(0.0534)	(0.0504)	(0.0540)
High/Low$_c$				0.671^a	0.342^a	0.106
				(0.111)	(0.104)	(0.112)
Market access$_c$				0.378^b	0.381^b	-0.317
				(0.191)	(0.180)	(0.193)
Port proximity$_c$				0.112^a	0.0310	0.0398
				(0.0282)	(0.0266)	(0.0285)
Diversity$_c$				-0.526	0.339	0.198
				(0.297)	(0.280)	(0.300)
N	257	257	257	257	257	257
R^2	0.49	0.36	0.40	0.60	0.43	0.55

Notes: See Appendix F, Table 8. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B.

Table 22 – First-stage regressions for IV estimates - Step 3, Low-skilled natives

	(1) IV1 Mig.	(2) IV1 Dens.	(3) IV1 Area	(4) IV2 Mig.	(5) IV2 Dens.	(6) IV2 Area
1964 Employment density$_c$	0.225^a	0.328^a	0.166^a	0.117	0.266^a	0.133^b
	(0.0584)	(0.0491)	(0.0532)	(0.0620)	(0.0546)	(0.0588)
1982 Land area$_c$	0.0155	-0.215^a	0.441^a	0.0155	-0.180^a	0.358^a
	(0.0481)	(0.0405)	(0.0438)	(0.0537)	(0.0472)	(0.0509)
1982 Agricultural share in employment$_c$	-0.106	0.125	-0.376^a	-0.0301	0.118	-0.270^a
	(0.0763)	(0.0643)	(0.0696)	(0.0808)	(0.0711)	(0.0766)
1990 Migrant/ Low-skilled$_c$	0.631^a	0.205^a	0.179^a	0.500^a	0.149^a	0.181^a
	(0.0543)	(0.0457)	(0.0494)	(0.0574)	(0.0505)	(0.0544)
High/Low$_c$				0.419^a	0.142	-0.0471
				(0.117)	(0.103)	(0.111)
Market access$_c$				0.691^a	0.592^a	-0.531^a
				(0.213)	(0.187)	(0.202)
Port proximity$_c$				0.112^a	0.0175	0.0490
				(0.0326)	(0.0287)	(0.0309)
Diversity$_c$				-0.414	-0.145	0.761^a
				(0.290)	(0.255)	(0.275)
N	257	257	257	257	257	257
R^2	0.41	0.34	0.40	0.49	0.38	0.43

Notes: See Appendix F, Table 8. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B.

Table 23 – First-stage regressions for IV estimates - Step 3, Rural migrants

	(1) IV1 Mig.	(2) IV1 Dens.	(3) IV1 Area	(4) IV2 Mig.	(5) IV2 Dens.	(6) IV2 Area
1964 Employment density$_c$	0.255^a	0.183^a	0.0968^b	0.187^a	0.203^a	0.0737
	(0.0535)	(0.0543)	(0.0486)	(0.0546)	(0.0568)	(0.0520)
1982 Land area$_c$	-0.0748	-0.317^a	0.422^a	-0.0831^b	-0.383^a	0.347^a
	(0.0407)	(0.0413)	(0.0370)	(0.0417)	(0.0434)	(0.0397)
1982 Agricultural share in employment$_c$	0.153^b	0.0942	-0.490^a	0.149^b	0.204^a	-0.318^a
	(0.0644)	(0.0654)	(0.0586)	(0.0708)	(0.0736)	(0.0674)
1990 Migrant/ Low-skilled$_c$	0.713^a	0.259^a	0.308^a	0.502^a	0.0467	0.339^a
	(0.0399)	(0.0405)	(0.0363)	(0.0490)	(0.0509)	(0.0466)
High/Low$_c$				0.429^a	0.603^a	-0.0787
				(0.0967)	(0.101)	(0.0920)
Market access$_c$				0.623^a	0.244	-0.506^a
				(0.176)	(0.183)	(0.167)
Port proximity$_c$				0.120^a	0.0863^a	0.0561^b
				(0.0261)	(0.0272)	(0.0249)
Diversity$_c$				-1.226^a	-1.013^a	1.031^a
				(0.228)	(0.237)	(0.217)
N	257	257	257	257	257	257
R^2	0.71	0.49	0.50	0.77	0.57	0.56

Notes: See Appendix F, Table 8. Standard errors in parentheses. Significance: ^{a}: p<0.01, ^{b}: p<0.05. Variable definition: See section 2 and Appendix B.