Migration Externalities in Chinese Cities

Pierre-Philippe Combes, Sylvie Démurger, Shi Li

SciencesPo
DEPARTMENT OF ECONOMICS

Discussion paper 2015-05

Migration Externalities in Chinese Cities

Pierre-Philippe Combes
Sylvie Démurger
Shi Li

Sciences Po Economics Discussion Papers

Migration Externalities in Chinese Cities*

Pierre-Philippe Combes† Sylvie Démurger‡ Shi Li§

March 3, 2015

Abstract

We analyse the impact of internal migration in China on natives' labour market outcomes. We find evidence of a large positive correlation of the city share of migrants with natives' wages. Using different sets of control variables and instruments suggests that the effect is causal. The large total migrant impact (+10% when one moves from the first to the third quartile of the migrant variable distribution) arises from gains due to complementarity with natives in the production function (+6.4%), and from gains due to agglomeration economies (+3.3%). Finally, we find some evidence of a stronger effect for skilled natives than for unskilled, as expected from theory. Overall, our findings support large nominal wage gains that can be expected from further migration and urbanisation in China.

JEL Codes: O18, J61, R23, J31, O53.

Keywords: migration; urban development; agglomeration economies; wage disparities; China.

---

*We are very grateful to Nancy Qian for sharing her historical data. We thank Theo Eicher, Florian Mayneris, Xin Meng, Mark Partridge, Sandra Poncet, Jacques-François Thisse and referees for their useful comments on earlier versions of the paper. Any remaining mistakes are our own. This paper is part of the Agence Nationale de la Recherche (ANR) research programs ANR-11-BSH1-0014 and ANR-12-GLOB-0005, whose financial support is gratefully acknowledged. Combes and Démurger are researchers employed by the French Centre National de la Recherche Scientifique (CNRS).

†Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, 2, Rue de la Charité, 13002 Marseille, France; Sciences Po, Department of Economics, 28, Rue des Saints-Pères, 75007 Paris, France. Also affiliated to the CEPR. Email: ppcombes@gmail.com.

‡Université de Lyon, 69007 Lyon, France; CNRS, GATE Lyon Saint-Etienne, 93 Chemin des Mouilles, 69130 Ecully, France; IZA, Bonn, Germany. Email: demurger@gate.cnrs.fr.

§School of Business, Beijing Normal University, China; IZA, Bonn, Germany. Email: lishi@bnu.edu.cn.

Introduction

In recent decades, China has witnessed a massive internal transfer of labour. By fuelling Chinese cities with millions of non-local workers, the so-called "Great Migration" plays a crucial role in the urbanization process of the country. The vast majority of migrant workers come from rural areas, although urban-to-urban migration also exists. Rural migrant workers - defined as agricultural people working and living outside the township of their household registration for more than 6 months - were estimated at 153 million in 2010, a rough doubling over the decade. As of 2010, they accounted for about 70% of the "floating" population[1] and 44% of urban employment (Chan, 2012).[2]

Internal migration is the main force of incremental urban population since the end of the 1990s. Yet migrant workers face major inequalities in Chinese cities, primarily because they are denied the status of city dwellers as long as they do not hold a local household registration or *Hukou*. Established in 1958, the *Hukou* system assigns every Chinese citizen a registration status (*Hukou*) that records not only their place of residence but also their type of registration. The *Hukou* type classifies people into "agricultural" or "non-agricultural", a classification which is usually referred to as "rural" and "urban" *Hukou*. By granting differential access to welfare benefits between urban and rural residents and between local and non-local residents, and by requiring official approval for a change in *Hukou*, the system has strongly restricted population mobility for decades. Nowadays, most migrants, especially those holding a rural registration, still hold their hometown *Hukou* and remain marginalized citizens in cities. In particular, they are denied equal access to public services; they face poor and unsafe working conditions; and they work primarily in the informal sector (Cai et al., 2008; Démurger, Gurgand, Li and Yue, 2009).

This article estimates the impact of migrants on local residents' nominal wages. As the migrant population is getting increasingly larger both in number and as a share of the local resident population, and despite the concentration of migrants in some specific segments of local labour markets, concerns have been raised regarding their potentially negative impact on local residents' labour market outcomes. This fear mostly concerns rural migrants, who are significantly less educated than their urban counterparts (Démurger et al., 2009; Deng and Li, 2010). By creating a large unskilled labour supply shock, the inflow of rural migrants could crowd out job opportunities for local workers and exert a downward pressure on wages.

[1] The floating population comprises all people with no local residency rights. It includes both migrant workers and their dependants (non-working people), originating from either rural or urban areas. In contrast, it excludes people who obtained a change in their household registration.

[2] The share of migrants in local employment sharply increased over the last two decades in China. Using the 2000 Census data, Cai, Park and Zhao (2008) estimate that migrants, both rural and urban, accounted for 19.6% of the employment in China's cities (excluding townships) in 2000.

Such concerns mirror similar issues raised in developed countries regarding international migrants. In their thorough review of the large and much-debated literature on the effects of immigration on local economies, Lewis and Peri (2015) conclude that immigration has a positive impact on local wages for most native-born workers in industrialized countries. In the case of U.S. immigration, earlier studies found dominant adverse effects on the economic status of the least-skilled competing native workers (Borjas, Freeman and Katz, 1997; Borjas, 2003) but more recent contributions show that the impact depends on immigrants' and natives' respective skill distribution and how native workers move across skill cells in response to immigration (Card, 2005; Ottaviano and Peri, 2012; Peri, 2012). Immigrants typically differ from native workers in terms of age, education and occupation. As summarized by Lewis and Peri (2015), these differences in skill characteristics have a number of theoretical implications that must be taken into account when estimating the wage effect of immigration. First, the effect depends on the degree of substitutability or complementarity among immigrant and native workers, which itself depends on natives' characteristics. Second, when confronted with immigration, native workers may chose to move either across space or across occupations in response to increased competition, and this mobility may mitigate the potential adverse effect of immigration. We show for China that migrants do exert quite a large positive effect on local residents' wages, be they skilled or unskilled, and that the effect is more positive for skilled workers, as expected from theory.

Most empirical studies focus on international migration in OECD countries (Docquier, Ozden and Peri, 2014), and systematic evidence is missing on how migration affects local productivity and wages in large developing countries, including China at the forefront, but also India or Indonesia. Moreover, the change of geographical scale for such countries and the relative smaller share of international migrants with respect to national ones makes it important to assess the role of internal migration. This paper contributes to filling the gap by assessing the extent to which internal migration in China affects urban residents' wages. Besides the effects identified in the literature on international migration, the simple increase in urban population induced by migration flows may also be beneficial to labour productivity through standard agglomeration economies.[3] A further contribution of this paper is to propose an empirical strategy to decompose the migrants' impact on local urban workers into the part due to their substitutability/complementarity in the production function and the part due to agglomeration effects.

[3]The standard explanation provided by economic geography for the positive impact of urban scale on productivity consists in a series of agglomeration effects that include pure knowledge spillovers, the sharing of inputs, and the pooling of the labour force. See Duranton and Puga (2004) for a thorough review of the micro-foundations of these effects and Combes and Gobillon (2015) for a survey of the empirics of agglomeration economies.

We proceed in two steps. We first introduce city fixed effects in a standard individual wage equation and assess the relative explanatory power of individual, firm and city effects. Among other things, we highlight the large role of city effects in China. In a second step, we explain the city fixed effects by both the share of migrants in local employment and a number of city variables capturing agglomeration economies. As detailed in Lewis and Peri (2015) and in Combes and Gobillon (2015), the estimation of both the migrant impact and agglomeration effects on individual wages raises a number of specific methodological issues, which we consider here. First, workers may sort spatially according to personal characteristics (e.g., their skills), which generally affects their labour outcome. If this is the case, it may be difficult to identify separately the role of local effects and the role of the workers' characteristics on their wages. We present evidence of a quasi-absence of local resident workers' sorting according to their observed skills across Chinese cities. Sorting on unobservable variables not correlated to observables remains a possibility, but it is highly unlikely in a country where mobility has been strongly restricted for decades. Second, and most importantly for measuring migration externalities, the estimation of the impact of local variables is almost inevitably plagued by a reverse causality bias. If large locations where the share of migrants is high enjoy higher productivity, and therefore higher nominal wages, then higher wages should also attract more workers, especially when they are mobile, thus reversing the causality. We attempt to estimate the causal impact of migrants through a thorough instrumentation strategy detailed in section 2.2.

There is a large and growing empirical literature analysing the Great Migration in China and its consequences on labour market outcomes for migrants or for their family members left behind. However, very few contributions relate to our approach. Meng and Zhang (2010) is the only paper we found that tries to specifically evaluate the impact of rural migrants on natives' labour market outcomes in China. They find no significant impact of the rural migrant inflow on average wages of urban workers, which may suggest that complementarities between native and migrant workers compensate for the negative effect of the increase in (unskilled) labour supply. However, this paper does not account for agglomeration effects caused by the inflow of migrants as we do here, and does not emphasise the geographical dimension of local labour markets that might be of particular importance in a fast urbanizing country.[4] In what follows, we propose an original approach to explore the simultaneous

[4] de Sousa and Poncet (2011) is another attempt to study the impact of migration on wages, in the manufacturing sector only. However, they use the flow of non-residents newly-arrived in a province (in the last 6 months) as a measure of migration, rather than the stock of migrants as we do here. Focusing on the flow of new migrants strongly limits the scope of the analysis because the largest part of the migrant population - those who have been in cities for more than 6 months - is simply missed. Moreover, the analysis is based on aggregate rather than individual data, on a very aggregated spatial level of 29 provinces only. It relies on a new economic geography model, which does not allow full separation of the migrant and agglomeration

impact of individual characteristics, geography, and migration in an urban perspective.

Our empirical analysis is based on individual data extracted from the 2007 Urban Household Survey conducted by the National Bureau of Statistics of China. The database is a representative sample of China's main regions, but the sampling frame includes only registered urban residents. In other words, migrants whose *Hukou* is not a local one are excluded from the sample. In order to identify migration externalities on urban residents' wages (our "natives" here), we use the 1% Population Survey of China conducted in 2005 to compute the share of the migrant population in cities. We find evidence of a positive and significant impact of the local share of migrants on natives' wages. Our estimates suggest that if new migrants move to a city at a constant number of local residents and if this move increases the migrant share in total employment from the first quartile (decile respectively) of the distribution of the migrant variable across Chinese cities to the last quartile (decile respectively), then the associated natives' wage increase in the city is 10% (33% respectively). Around two-thirds of that figure comes from the complementarity of migrants with natives, and a third results from agglomeration effects induced by the increase in total employment density generated by the migrant inflow. If migrants were to replace natives in local employment, thus keeping total employment density constant, the productivity of natives would increase by 6.4% (20.6% respectively) just through the complementarity effect. We also find evidence of a 25% lower total impact (50% lower for the complementarity effect) for unskilled natives compared with skilled ones, though the differences are only marginally significant because of insufficiently precise estimates. We control for some agglomeration effects other than density and we instrument the a priori most endogenous variables. Over-identification and weak-instruments tests are performed, and passed, for different sets of instruments. However, omitted permanent characteristics of cities relating to first-nature geography, transport infrastructure or institution quality, for instance, might be correlated with wages, migrant and density variables, and instruments. Further studies would be required to fully confirm the magnitudes we obtain here and the causal impact of the migration and density variables.

The rest of the paper is organised as follows. Section 1 presents the empirical strategy followed in the paper. Section 2 describes the data sources and justifies our instrumental variables strategy. Section 3 compares OLS and various IV estimates and discusses the estimated magnitude of the migrant effects in Chinese cities. Section 4 concludes.

effects. A moderate downward pressure of the migrant variable on manufacturing wages is obtained when all variables are instrumented by GMM.

1 Empirical strategy

This section presents the framework for the simultaneous identification of substitutability / complementarity and agglomeration effects of migrants. It also discusses possible endogeneity biases arising from missing variables and reverse causality.

1.1 Substitutability/complementarity *versus* agglomeration effects of migrants

The literature on immigration emphasises that migrants can have either a positive impact on natives' labour market outcomes, when their skills complement those of natives, or a negative one when they substitute for natives. Typically, the wage of a native worker i, w_i, is regressed on a vector of individual characteristics X_i and the share of migrants in the total employment of city c, L_c^T, which is given by $MigSh_c = \frac{L_c^M}{L_c^T}$, where L_c^M is migrant employment. One estimates:

$$\log w_i = X_i\, \alpha + \lambda\, MigSh_c + \varepsilon_i. \tag{1}$$

where ε_i is a generic random component. A positive λ illustrates the case where complementarities between natives and migrants increase the wages of native workers when the local share of migrants increases. Importantly, Lewis and Peri (2015) emphasise that λ corresponds only to the partial effect of migrants, which is evaluated for a given occupation and location choice of the natives. As migrants affect local wages, this should in turn impact occupation and location choice, which implies a feedback effect on natives' wages, not captured by λ. In the Chinese context, where occupational choice by local residents and migrants is segregated and mobility is constrained, we believe that this effect should not cause any great difference between partial and total migrant effects. In any case, the cross-sectional nature of available data does not allow us to investigate individual occupation and location choice. Strictly speaking, we only identify the partial effect of migrants, as most of the early literature on international migration.

Lewis and Peri (2015) also emphasise that beyond controlling for natives' characteristics, it is interesting to investigate the role of migrants' characteristics as well. Then one can use such information to recover the average partial effect that corrects λ for the share of migrants in each type (in terms of education level or gender for instance). Again, data limitation prevents us from following such a route. Still, providing a first estimate of λ for China is important, especially since we control for another series of local effects not always considered in the immigration literature.

Indeed, one limit to specification (1) relates to the fact that it does not control for local effects that emerge from agglomeration economies.[5] The local economy characteristics - first and foremost its size - affect technological spillovers and pecuniary economies (better access to final and intermediate inputs), and therefore local productivity. In turn, this affects local nominal wages, with a possibly large impact[6]. Let $Den_c^T = \frac{L_c^T}{Area_c}$ denote total employment density in city c, with $Area_c$ being the land area of city c. Typically, the agglomeration literature estimates:

$$\log w_i = X_i\,\alpha + \beta\, Den_c^T + \varepsilon_i. \tag{2}$$

A positive β reflects the fact that denser areas induce agglomeration gains, which translate into higher nominal wages. Because agglomeration also generates congestion effects, β only reflects the total net effect of gains and costs from density. Typically, it may be found negative when cities are too large. We discuss below the role of other agglomeration variables, capturing in particular the absolute city size and not just density.

Specifications (1) and (2) are not consistent with each other, because each one misses the other's effects. When panel data are available, some authors in the immigration literature control for local fixed effects. However, this strategy does not assess the role of time-varying agglomeration economies. Neither does it consider that the impact of the migrant share is a combination of the substitutability/complementarity of migrants and of agglomeration effects. Indeed, the arrival of migrants in a city also impacts natives' outcomes because migrants make the city larger. This is one of our contributions, in proposing a specification that encompasses both types of effects and allows us to identify each channel separately. To see this, let us first note that the total employment density can be decomposed as follows:

$$Den_c^T = \frac{L_c^N + L_c^M}{Area_c} = \frac{Den_c^N}{1 - MigSh_c} \tag{3}$$

where L_c^N and L_c^M are native and migrant employment in city c respectively, and Den_c^N is the local density of native employment. Equation (2) leads to:

$$\log w_i = X_i\,\alpha + \beta \log Den_c^N + \beta \log\left(\frac{1}{1 - MigSh_c}\right) + \varepsilon_i. \tag{4}$$

This illustrates the idea that, for a given number of local natives, an increase in the num-

[5] Borjas (2003) advocates controlling for the size of the native workforce in order to provide a correct interpretation of the impact of an increase in the size of the immigrant population on native labour market outcomes. This could control for part of the agglomeration effects, but we push the intuition further here.

[6] See Combes and Gobillon (2015) for a review of theoretical underpinning, estimation strategy, and standard estimates for such agglomeration effects.

ber of migrants, and therefore of the migrant share, increases natives' wages through agglomeration economies reflected in a positive β. But (4) does not encompass the substitution/complementarity effects between migrants and natives that enter (1). One can augment (4) to consider both, which gives:

$$\log w_i = X_i\,\alpha + \beta \log Den_c^N + \beta \log\left(\frac{1}{1 - MigSh_c}\right) + \lambda MigSh_c + \varepsilon_i. \tag{5}$$

This specification, estimated under the constraint that the effects of the density of natives and of the first variable for the migrants' share are identical, would allow us to identify both the agglomeration and the substitution/complementarity effects of migrants separately. However, it would be highly dependent on the parametric choices made, and its estimation would probably not be very robust, as the two migrant variables are likely to be highly correlated (they are almost identical when $MigSh_c$ is small). Furthermore, specification (5) also underlines the role of the specification adopted for the substitution/complementarity migrant effect in (1). The specification of the substitution/complementarity effect depends on the functional form chosen for the production function, and the way native and migrant labour enter it, which is an ad hoc choice. If we assume that the substitution/complementarity migrant effect is proportional to $\log\left(\frac{1}{1-MigSh_c}\right)$, instead of $MigSh_c$, we obtain:

$$\log w_i = X_i\,\alpha + \beta \log Den_c^N + (\beta + \widetilde{\lambda}) \log\left(\frac{1}{1 - MigSh_c}\right) + \varepsilon_i, \tag{6}$$

where $\widetilde{\lambda}$ is the new migrant substitutability/complementarity parameter. Specification (6) uses the impact of the employment density of natives to identify the agglomeration effect, and the total effect of the migrant variable corresponds to the sum of the migrant agglomeration effect and the migrant substitution/complementarity effect.

$\widetilde{\lambda}$ can be obtained by subtracting the estimate for β from the estimate for $\beta + \widetilde{\lambda}$. However, such a separate identification of each effect is only possible when agglomeration effects from natives and migrants are assumed to be identical, as implicitly assumed in equation (2), which considers total employment. If this were not the case, one would only identify the total (agglomeration plus substitution/complementarity) effect of migrants but, crucially, agglomeration effects from natives would remain to be controlled for in the specification. This is important because natives' density and migrants' share can be correlated in data sets, and not controlling for the former, as often done in the immigration literature, would bias the estimated impact of migrants.

The following specification could also be estimated:

$$\log w_i = X_i\,\alpha + \beta \log Den_c^T + \lambda MigSh_c + \varepsilon_i. \tag{7}$$

However, this specification, only valid under the assumption of identical agglomeration effects for natives and migrants, is also difficult to interpret because the total density is itself a function of the share of migrants. Moreover, in the Chinese case, raw data correspond to the employment density of natives only, and total employment density is obtained by adding migrant employment that is less precisely measured. Both explanatory variables are subject to measurement errors in (7), whereas this is only the case for one in (6). Similarly, reverse causality mostly affects migrant employment, and therefore affects both variables in (7) but only one in (6). As a consequence, equation (6) should also reduce endogeneity issues.

1.2 Other local effects and endogeneity concerns

Apart from employment density, other agglomeration economies variables are typically included as additional explanatory variables in the economic geography literature. This is the case for the city land area. Whereas density captures the role of the city size in terms of thickness, land area captures its spatial extent. When agglomeration gains dominate, both are expected to positively affect wages. Moreover, following Jacobs (1969), one may also expect a diverse local economic environment to favour innovation and productivity, and thereby wages. This intuition is usually captured by introducing the inverse of a sector concentration Herfindhal index, which is what we do here. The city location within the network of all other cities may also matter. Typically, if workers do benefit from the density of their own city, they should also benefit, although probably to a lower extent, from the density of neighbouring cities with which they are likely to interact. Economic geography models have considered such a role of the access to distant markets. Empirically, it is usually captured by market potential variables. We use the Harris (1954) definition, which simply corresponds to the inverse-distance weighted sum of densities over all Chinese cities other than the city considered. Following Au and Henderson (2006), we add as another control for the proximity to external markets the distance to the closest seaport,[7] which reflects the access to foreign markets.[8] Last, agglomeration economy variables also typically include the role of the local

[7]Among the 12 major coastal seaports in China (in terms of the volume of freight handled), which are, in decreasing order: Shanghai, Ningbo, Guangzhou, Tianjin, Qingdao, Qinhuangdao, Dalian, Rizhao, Yingkou, Yantai, Lianyungang and Zhanjiang.

[8]Some studies, such as Hering and Poncet (2010) for China, use a structural version of market potential, where a single variable encompasses the role of the city's own size, the proximity to other large Chinese cities, and the proximity to foreign markets. The specification we choose here is more flexible since it identifies a

sector size, or specialisation, measured by the share in the local economy of employment in the firm's sector. Some technological spillovers or pecuniary economies can relate more to the size of the sector within the local economy than to the overall size of the local economy.

Importantly, in order to simplify the treatment of heteroscedasticity and endogeneity, the literature usually estimates the role of local variables in two steps (Combes and Gobillon, 2015):

$$\log w_i = X_i\,\alpha + L_{c(i)s(j(i))}\,\gamma + \delta_{c(i)} + \varepsilon_i \tag{8}$$

$$\delta_c = (\beta + \widetilde{\lambda})\log\left(\frac{1}{1 - MigSh_c}\right) + \beta\log Den_c^N + U_c\,\eta + \nu_c \tag{9}$$

where $\delta_{c(i)}$ is a fixed effect for the city c where individual i locates. It captures the role of any city effect, observed or not, that does not depend on the individual's sector. The vector of individual characteristics that affect productivity, X_i, includes gender, experience, education and occupation.[9] Some authors also use certain characteristics of the firm in which individual i is employed. We consider here sectoral dummies, which allows us to focus only on the spatial, and not the industrial, variations in wages. In the specific case of China, the ownership of firms is another important determinant of workers' wages (Chen, Démurger and Fournier, 2005). Local effects that also depend on the sector, such as specialisation, L_{cs}, are directly introduced in the first step as well. Then, in the second step, the city fixed effect is regressed on the migrant variable, controlled for the role of the employment density of natives, as suggested by (6), which is augmented by additional city variables embodied in vector U_c. This second step is instrumented separately from the first step estimation.

OLS estimates for specification (9) can clearly suffer from endogeneity bias. Typically, migrants are attracted by higher expected wages in the city, which they probably infer from the high wages of natives (i.e., a high δ_c). This means that the share of migrants in the city is affected by reverse causality. OLS estimates of agglomeration effects may also be biased because of missing productive amenities in the wage specification. For instance, dense areas may benefit from better public infrastructure (e.g., large train stations, airports or universities). By making labour more productive, such amenities increase wages. They also attract migrants to the city, who directly consume these amenities. This causes both density and the share of migrants in the city to be correlated with the second step random component.[10]

separate effect for these three underlying components.

[9]Within-the-firm experience is also sometimes considered but, unfortunately, such information is not reported in the data set we use.

[10]Conversely, there is no reason to control for a local cost of living/price index. This point, made several times in the literature, results from the fact that the estimated specification is derived from the firms' optimal use of labour, and not from a model of employees' location choices (see Combes and Gobillon, 2015, for more

We tackle both endogeneity concerns by instrumentation and we instrument both the migrant share and the native density variables, although endogeneity concerns are stronger for the former. Indeed, they arise for migrants from both reverse causality and missing amenities, and, most importantly, for recent location choices, whereas the population we refer to as natives has either never moved or migrated long ago. In some specifications we also extend instrumentation to the market potential variable, as it depends on native employment in nearby cities, which may be correlated with own city's native employment. One could argue that land area and industrial diversity could also be affected by reverse causality, but here the links start to become pretty weak and instrumenting for so many variables at the same time is generally not meaningful. Still, we present estimations where land area is also instrumented, and we verify that not controlling for other non-instrumented agglomeration variables does not lead to significantly different estimates. For any further specification and interpretation concerns of this approach and the solutions to deal with endogeneity, which are now standard in economic geography, one can refer to Combes and Gobillon (2015).

2 Data and instruments

2.1 Data sources

The data used are drawn from three main complementary sources: 1) individual data for urban residents extracted from the 2007 Urban Household Survey conducted by the National Bureau of Statistics of China (NBS), 2) migrant share in cities extracted from the 1% 2005 China Population survey, and 3) city-level data compiled from the 2008 *China City Statistical Yearbook* published by the NBS. Additional data for instruments come from historical maps and from Au and Henderson (2006).

The raw individual database is a sub-sample of the larger NBS Urban Household Survey. It includes a subset of 16 provinces[11] selected to obtain a representative sample of China's four major regions (large metropolitan cities, coastal region, central region and western region). The 87 sampled prefecture-level cities[12] were chosen following a stratified random sampling method, and 10,000 households within cities were randomly sampled.

details on this issue).

[11]Beijing, Shanxi, Liaoning, Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Henan, Hubei, Hunan, Guangdong, Chongqing, Sichuan, Yunnan and Gansu.

[12]The NBS survey also covers four autonomous prefectures, all in Yunnan province. However, we cannot keep them in the analysis because we do not have any information at the city-level such as area or total employment for these autonomous prefectures.

A major feature of the sample is that it only contains registered urban residents (i.e., the urban *Hukou* holders). This means that the observed wage disparities we focus on only concern urban residents ('natives'). Migrants are accounted for through their share in city employment, which is calculated from the 1% Population survey issued in 2005. Migrants are defined as working individuals aged 16 to 60 who are not living in their county of *Hukou* registration.[13] Importantly, this includes migrants from both rural areas and other urban areas.

The sample is restricted to individuals aged 16 to 70, who declared working at least part of the year and earning (positive) wages. Owners of private or individual enterprises are excluded because we cannot separate wages from profit in their case. Taking into account these restrictions, we are left with 14,590 urban resident workers. The earnings variable is the declared income from wage employment, which includes the basic salary and all sorts of bonuses, allowances and subsidies (including housing or medical subsidies), other wages (including overtime wages), and other income from work unit. A major drawback to the NBS data is that no working times are recorded, which does not allow us either to separate part-time and full-time workers, or to estimate an hourly wage model. Given this constraint, we focus on total earnings from the work unit reported by wage workers who declared strictly positive annual earnings.

City-level data come from the *China City Statistical Yearbook* 2008. The administrative structure in China comprises four levels. In 2007, the mainland territory was divided into 31 province-level regions (excluding the two Special Administrative Regions, Hong Kong and Macao), 333 prefecture-level regions (including 286 prefecture-level cities), 2,859 county-level regions and 40,813 township-level regions (National Bureau of Statistics, 2008). Prefecture-level cities in China typically include a city, its suburban districts, and a number of rural counties under the jurisdiction of the city government. Employment and area statistics are available at two levels: the whole prefecture and the metropolitan area (composed of the city and suburban districts only). All the city-level variables used in the analysis refer to the second level, the metropolitan area. Kernel density estimations, available upon request, for the distribution of employment density, respectively for all prefecture-level cities in China and for the 87 prefecture-level cities of our sample, show fairly close distribution patterns, and confirm the representativeness of our sample of cities.

Table 4 in Appendix A reports summary statistics for all the variables related to migrant and agglomeration effects. Spatial variations of our two main variables of interest, the density of natives and the share of migrants, are quite large. Density at the last decile is almost 40

[13]The duration of migration can also be added to the definition of a migrant. The standard definition of the NBS only considers individuals who have resided in the city for at least 6 months. However, using this definition does not affect our conclusions.

times larger than at the first decile, which is even larger than what is usually observed for developed countries, thus making it important to control for in the estimations of the migrant impact. For the migrant variable, the ratio is over 10. The raw migrant share varies from 8.4% at the first decile to 62.4% at the last decile. Other local variables present significant spatial variations as well, but to a lower extent for some of them.

2.2 Instruments

To address endogeneity, we take an instrumental variable approach and we use several sets of instruments as sources of exogenous variation for the suspected endogenous variable(s). The standard practice since Ciccone and Hall (1996) to instrument for urban size or density is to use historical variables such as long lags of population density. The rationale is that the spatial distribution of population is persistent over time but the sources of productivity differences, and therefore of wage differences, differ over time. Large places that were already focal points for migrants centuries ago are still attractive today and act as a magnet for modern migrants without necessarily being the places where wages currently benefit from the largest shocks. Following this intuition, our first series of instruments relates to historical data. As argued by Wu and Gaubatz (2013), the Chinese urban system is one of the world's oldest. It has evolved over time with major historical events, including the establishment of foreign treaty ports that were part of the 'unequal treaties' that China signed with Western countries in the late Qing dynasty.[14] Although we do not have historical population data, we identify important cities in the late 19th and early 20th centuries that became major economic centres after the opening of the 48 treaty ports conceded to foreign countries between 1842 and 1920.[15] From these historical data, we compute several indicators. First, we use a dummy variable for cities that are either historic major cities or former treaty ports. Second, we compute a 'peripherality' index (average distance away) to these historic cities. This captures the intuition that proximity to historical cities is a substitute from being exactly located there, since interactions remain possible at not too long distances. In the same spirit, we also consider an overall peripherality index that consists in the average distance of any city to all cities, be they historic or not. Economic geography shows that "being central" in the economy influences productivity, wages, and employment. Theory also suggests that distance should be weighted by the level of economic activity when measuring centrality, for the same reason that both local density and market potential variables are included in the specification. This is also the reason why our peripherality instruments are not weighted by population or employment. Therefore, they remain correlated with instrumented variables

[14] The system was abolished in 1943 after China signed new treaties with Britain and the US (Jia, 2014).
[15] Nancy Qian generously provided her data on these historical cities.

but the unweighted measure increases the probability of not being too correlated with current wage shocks.

The second set of instruments is borrowed from Au and Henderson (2006), who use 1990 data on city characteristics and amenities. We use the share of the manufacturing sector in total employment, the share of non-agricultural employment in total employment, doctors per capita and telephones per capita, all measured for the year 1990. The intuition is that the past industrial composition of the cities should have influenced their total employment growth. Similarly, health or communication amenities that are valued by consumers constitute standard determinants of migration flows. On the other hand, with an almost twenty-year lag over a period of major reforms in China, past industrial composition, and even more obviously past doctors and telephones rates, should not be too correlated with current wage shocks.

Finally, the third set of instruments is composed of an instrument in the spirit of Bartik (1991) but computed with cross-section data. The index corresponds to the population that the city would have, if its workers were located in a city with total employment corresponding to the mean city employment (nation-wide) for workers in the same occupation. More precisely, to build this "occupation Henderson instrument" (as defined by Combes, Duranton and Gobillon, 2012), we first compute the mean city employment for each occupation. Then for each city, we compute the instrument by interacting the local share of employment for an occupation with the mean employment at the city level for this occupation and then summing across all occupations. Put differently, a city with a high proportion of managers will be predicted to be large because on average over China, managers locate in larger cities, while an urban area with a high proportion of blue-collar occupations will be predicted to be small. This sort of instrument is interesting because it removes from the city size the part that is not explained by its occupational structure, typically the part that could relate to possibly missing variables in the wage equation and that makes the city actually larger or smaller, or attractive for migrants. Therefore, such an instrument should be fairly exogenous to the random wage component and simultaneously be strong enough because the occupational employment structure necessarily affects at least part of the city size.

Using instruments from different families should reduce the risk of weak instrument issues and give more credence to over-identification tests. The cost could be some large differences in estimates when using the different sets of instruments, because only local average treatment effects (LATE) are identified, but this does not seem to be the case here, since all IV estimations lead to very consistent estimates across specifications.

3 Results

3.1 The explanatory power of individual wages and local effects

Before moving to our main second step estimations, corresponding to specification (9), we provide a variance analysis of the first step estimation (specification 8) in order to assess the main sources of variation in Chinese data. We estimate first step regressions that successively include different sets of explanatory variables: location effects, individual characteristics, and firm characteristics. As our interest here is in the relative contribution of the different sets, we do not report the full estimation results but only focus on their adjusted R^2, which are displayed in Table 1, and on a variance analysis of the complete specification, which is reported in Table 5 in Appendix B.

As expected, individual characteristics alone explain the largest share (25%) of the variations in individual wages, and the explanatory power of firm characteristics alone is smaller, at 13%. Much less documented in the case of China, we also find that location accounts for a non-negligible share of the variations in individual wages: city dummies and specialisation together exhibit an explanatory power of 17%. Table 5 in Appendix B confirms that individual characteristics are the main factors explaining individual wage disparities, but that location effects matter and come in second place, above firm effects. Table 5 also indicates that among local effects, specialisation explains very little of wage disparities, since local effects operate at the global city level, not within sectors. These results closely match those obtained for developed countries (Combes and Gobillon, 2015).

Table 1: Explanatory power of various sets of right-hand side variables

Adj. R^2 for individual wages in 2007 (log wage) as a function of:	
City effects	0.17
Individual characteristics	0.25
Firm characteristics	0.13
City effects and Individual characteristics	0.41
City effects and Firm characteristics	0.32
Individual characteristics and Firm characteristics	0.28
All three sets	0.44
N	14,590

Notes: City effects include both city fixed effects and the specialisation variable.

Importantly, Table 1 also reveals a fairly strong orthogonality of the three groups of effects. City effects and individual characteristics together explain 41% of wage disparities when the sum of their individual R^2 is 0.42, while city and firm effects explain 30% for a sum of their individual R^2 at 0.28. Individual and firm effects are slightly more correlated. This means that higher wages observed in some cities cannot be attributed to differences in the composition of the labour force or the type of firms present there. Put differently, there is no sorting of workers across cities according to their skills, a finding corroborated by the absence of correlation between the effect of individual characteristics and city-level dummies (with a non-significant correlation coefficient at 0.0012). This finding for China is in sharp contrast with what is usually obtained for developed countries, where a large fraction of the explanatory power of city effects arises from the sorting of workers (see Combes and Gobillon, 2015).

To further examine the absence of sorting hypothesis, we run additional regressions not reported here. If individuals were sorted across cities according to their abilities, some local variables - density and the migrant share in particular - would be correlated with individual characteristics such as education or occupation. Hence, one would expect their estimated effect in the second step to change when individual characteristics are not included in the first step specification. As for China, the elasticity of the density and of the migrant effect is found to be very stable across specifications, with and without individual controls in the first step.

Sorting could occur on unobserved characteristics too. Combes, Duranton and Gobillon (2008), who control for a worker fixed effect in a panel setting for France, show that individual fixed effects are actually strongly correlated with density. Not controlling for them overestimates the density effect by a factor of 2. The cross-section nature of our data prevents us from implementing a similar strategy. This should not be a major concern though, since we show that even observed individual characteristics such as education or firm ownership are not correlated with local effects, and no sorting on those observed characteristics takes place. It is difficult to imagine unobserved individual characteristics that would not be correlated with our individual observed variables and yet would be correlated with the migrant and density variables. Therefore we are confident that no large bias emerges from individual sorting. Since our database covers urban *Hukou* holders only, the sorting hypothesis refers in any case to this part of the urban population alone. In this context, the absence of sorting is not surprising, because labour mobility towards urban areas with a change in *Hukou* has been highly restricted for decades and remains controlled. Still, individual sorting will deserve further investigation once suitable data are available for China.

3.2 The impact of migrants

Our main results are provided in Table 2,[16] which displays both OLS and IV estimations for the second step specification.[17] In the two-stage setting we adopt here, a concern regarding the second stage estimation arises because we use estimated fixed effects as the dependent variable. To account for the measurement error thus generated, we also compute feasible generalised least square (FGLS) estimates for the second step. As shown in Appendix D, no major difference with the OLS estimates presented in Table 2 emerges. This may be attributed to the precise estimation of local effects in the first stage, itself resulting from the use of individual data and the presence of marked spatial wage disparities in China.

Our first main result relates to the large univariate correlation between city fixed effects and the city share of migrants. Alone, the migrant variable explains 54% of the variance (column (1)) when employment density alone explains 18% (column (2)). The migrant variable in China presents a correlation with city fixed effects similar to the correlation of city density with city fixed effects in developed countries. Conversely, the correlation between city fixed effects and density appears to be much lower in China.

The migrant share and the employment density are pretty highly correlated, with a correlation at 0.34 (see Table 8, Appendix E). As explained in section 1.1, they must be introduced simultaneously for a correct interpretation. This is done in column (3). The explanatory power is only slightly above that of the migrant variable alone, and as expected, both estimated elasticities decrease in magnitude.

Although the specification with these two variables misses other agglomeration effects that we discussed in section 1.2, one can attempt to instrument it in order to estimate a more causal effect of city migrant share and density on natives' wages. The corresponding IV estimates are reported in column (4). Such instrumentation is demanding because of both reverse causality and many possibly missing effects, but we produce an estimation where the over-identification test is passed (with a Hansen p-value of 0.25), and instruments, based on peripherality, the number of doctors and telephones per capita, and the city size predicted from its occupational structure, pass weak instruments tests.[18] The impact of the migrant

[16] As some of the instruments are available for 83 cities only, the sample is reduced to these cities for the second step estimations. OLS estimates on the whole sample of 87 cities provide similar results.

[17] Table 6 in Appendix C displays estimation results for the first step specification. It fully confirms usual findings on the determinants of urban wage settings in China.

[18] The Cragg-Donald statistics we obtain are above the rule of thumb of 10 generally considered as sufficient for instruments not to be weak. More precisely, however, they are close to the 5% critical values proposed by Stock and Yogo (2005). In that case, Limited Information Maximum Likelihood (LIML) estimates provide more reliable point estimates and test statistics than Two Stage Least Squares (TSLS). Therefore we use LIML here. Cragg Donald values are now well above the Stock and Yogo (2005) weak instruments critical values for LIML estimates reported in the notes of Table 2. In any case, TSLS point estimates are in general only different from LIML ones at the second or third decimal place.

Table 2: The migration impact - OLS and IV estimates for the second step

	(1) OLS1	(2) OLS2	(3) OLS3	(4) IV1	(5) OLS4	(6) IV2	(7) IV3	(8) IV4	(9) IV5
Migrants	0.405***		0.369***	0.394***	0.279***	0.326***	0.324***	0.325***	0.322***
	(0.042)		(0.043)	(0.074)	(0.046)	(0.077)	(0.065)	(0.073)	(0.067)
Density		0.099***	0.046**	0.089***	0.067***	0.124***	0.105***	0.121***	0.105***
		(0.024)	(0.018)	(0.025)	(0.024)	(0.034)	(0.030)	(0.035)	(0.025)
Land area					0.103***	0.128***	0.118***	0.123***	0.119***
					(0.030)	(0.034)	(0.033)	(0.034)	(0.044)
Diversity					−0.014	−0.182	−0.136	−0.175	
					(0.135)	(0.153)	(0.142)	(0.147)	
Market potential					0.182*	0.050	0.081	0.010	
					(0.096)	(0.119)	(0.105)	(0.128)	
Distance to seaport					−0.005	0.005	0.002	0.002	
					(0.013)	(0.014)	(0.013)	(0.014)	
Instruments:									
Peripherality	N	N	N	Y	N	Y	N	Y	Y
Henderson occupations	N	N	N	Y	N	Y	N	N	Y
Doctors per capita	N	N	N	Y	N	Y	N	Y	N
Telephones per capita	N	N	N	Y	N	N	Y	Y	N
Historic city	N	N	N	N	N	N	Y	N	Y
Distance to historic c.	N	N	N	N	N	N	Y	N	Y
Manufacturing share	N	N	N	N	N	N	Y	N	Y
Non-agricultural share	N	N	N	N	N	N	Y	Y	Y
R^2	0.54	0.18	0.57		0.65				
Hansen p-value				0.25		0.11	0.66	0.17	0.59
Cragg-Donald				11.4		12.3	14.1	10.4	7.2
1st Shea part. R^2, mig.				0.37		0.38	0.49	0.42	0.43
1st part. Fisher, mig				13.7		10.8	14.2	13.1	9.5
1st Shea part. R^2, den.				0.56		0.58	0.60	0.49	0.62
1st part. Fisher, den				32.9		24.3	21.6	33.5	15.7
1st Shea part. R^2, mp.								0.58	
1st part. Fisher, mp								35.2	
1st Shea part. R^2, area									0.41
1st part. Fisher, area									7.3

Notes: 83 observations for each regression. The dependent variable is the city fixed effect estimated from equation (8). Estimation results for the first step are reported in Table 6, Appendix C. All regressors are logged. Standard errors in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$. In columns (4), (6) and (7), the migrant variable and density are instrumented. In column (8), the migrant variable, density and market potential are instrumented. In column (9), the migrant variable, density and land area are instrumented. In the instruments list, 'Y' means that the variable is used as an instrument, 'N' that it is not. LIML estimates for the parameters are reported. Critical values for the weak instrument test at the 5% significance level for a 0.10 desired maximal LIML size (most demanding size in Stock and Yogo, 2005) are 4.7 for IV1, 5.4 for IV2 and IV5, 4.3 for IV3, 8.7 for IV4. The first stage regressions for IV estimates are displayed in Tables 9 and 10, Appendix F.

share is slightly larger compared with the OLS estimates reported in column (3), but the increase is not significant. In contrast, the density elasticity increases more, to a level close to three times the standard values obtained for developed countries.

Since other agglomeration effects are not explicitly introduced, results from this estimation need to be confirmed, primarily because they are highly dependent on instrument validity. Moreover, they are slightly difficult to interpret because all agglomeration variables are intertwined and largely correlated with each other, and this is also the case with the migrant variable, as can be seen in Table 8 in Appendix E. If one wants to assess the migrant impact for given agglomeration effects, introducing all of them is the strategy to follow, and this is presented in column (5). First, introducing these additional agglomeration variables does indeed increase the explanatory power of the model, which reaches the level usually found for developed countries.[19] Second, the elasticity of the migrant variable and that of density are slightly lower. This is because most other agglomeration variables are positively correlated with them while also increasing wages.

Column (6) reports IV estimations, using the same instruments as in column (4) from which we remove the number of telephones per capita, which is not necessary to obtain a precise estimation. The instruments still pass the over-identification test, as expected since more control variables are introduced and one instrument is removed. Instruments also remain strong even if many control variables, also used as instruments, are introduced into the specification. Both the migrant and density elasticities increase, to a large extent for the latter. The migrant impact reaches a value at an intermediate level with respect to the estimates in previous columns.

Of the other agglomeration variables, only land area has a significant impact on the city fixed effect in column (6). Market potential, which has a slightly significant impact according to OLS estimates, is not significant any more. This is not due to a lack of precision or an increase in its standard error, but to a lower value of the estimated parameter itself. Therefore, in contrast to developed countries for which land area is rarely significant, the spatial extent of the city for a given density also increases local wages in China, on top of the positive impact of density on wages. Furthermore, access to other Chinese cities and foreign markets play no significant role, whereas they usually matter in developed economies. Diversity plays no role on wages in China as is often the case in developed countries.[20]

We finally run three further IV estimations to check the robustness of these conclusions. Considering different sets of instruments does not lead to different values here. This is

[19]One should bear in mind that in China, the explanatory power arises to a large extent from the migrant variable, not from employment density.

[20]The sector classification we use is not very detailed since we only have 11 sectors. As a consequence, our index may not properly account for the actual industrial diversification.

apparent when we use a set of instruments that does not intersect at all with the previous one, as reported in column (7). This set includes a dummy variable for being a historic city, the average distance to historic cities, the past city shares of manufacturing and non-agricultural employment, and the number of telephones per capita. These instruments also pass the over-identification test and are not weak.[21]

The literature usually considers that market potential also raises a reverse causality concern, and should be instrumented too, because workers' and firms' location choices are spatially correlated, as illustrated by economic geography models. Column (8) instruments migrant share, density and market potential simultaneously. Although such an instrumentation starts to be very demanding, a sub-set of our instruments satisfies over-identification and weak instruments tests. They lead to estimates very close to those of the two previous IV estimations.

Equally, one could argue that land area is also endogenous because the borders of cities are occasionally redefined according to population and employment growth. Instrumenting four variables at the same time is not reasonable. Therefore, the last IV estimation instruments for migrant share, density and land area, but does not consider the impact of other variables (column (9)). This reduces the possibility of bias due to the endogeneity of the control variables. Moreover, none of them was significant in previous estimations. A combination of our previous instruments passes over-identification tests, and they are only slightly weaker. Explaining land area is difficult because it rarely varies, for reasons not always directly related to our variables of interest. Still, under these quite different assumptions, elasticities for migrant share, density and land area are found to be very similar to the previous ones.

Overall, the different estimations confirm the robustness of the effects we estimate. The elasticity of the migrant share is around 0.32 in our most complete estimations, IV2 to IV5 in Table 2, that of employment density around 0.11, and that of land area 0.12. According to our discussion in section 1.1, the impact of the migrant variable is the sum of migrants' substitutability/complementarity and agglomeration effects, $\beta + \widetilde{\lambda}$. If one assumes that the agglomeration effect of migrants is similar to that of natives, migrants are complementary to natives in the production function, and the positive externality is estimated at around $0.32 - 0.11 = 0.21$. In other words, about two-thirds of the total impact of migrants on natives' wages results from migrant/native complementarity in the production function and the rest from agglomeration economies.[22]

[21] Note also that in general first stage estimations reported in Appendix F present expected signs for all our IV estimations, even if interpretation is not always easy given the number of instruments and control variables considered. In particular, both employment density and the migrant share increase in most cases with the city size predicted by occupations, the number of doctors or telephones per capita, the proximity to historical cities, or the manufacturing share.

[22] We also made some attempts to estimate specification (7), although we believe it to be less relevant for

Table 3: The migration impact - Skilled and Unskilled separately

	(1)	(2)	(3)	(4)
	Skilled		Unskilled	
	OLS1	IV1	OLS2	IV2
Migrants	0.295***	0.368***	0.257***	0.265***
	(0.048)	(0.075)	(0.054)	(0.085)
Density	0.067***	0.112***	0.067***	0.137***
	(0.025)	(0.036)	(0.027)	(0.040)
Land area	0.114***	0.126***	0.101***	0.133***
	(0.031)	(0.035)	(0.034)	(0.040)
Diversity	−0.057	−0.213	−0.011	−0.167
	(0.141)	(0.153)	(0.157)	(0.171)
Market potential	0.155	−0.004	0.262**	0.065
	(0.100)	(0.133)	(0.112)	(0.149)
Distance to seaport	−0.012	−0.005	0.005	0.011
	(0.014)	(0.015)	(0.015)	(0.016)
R^2	0.65		0.57	
Hansen p-value		0.33		0.15

Notes: 83 observations for each regression. All regressors are logged. Standard errors in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$. Instruments for columns (2) and (4) are identical to those in Table 2 column (8). Weak instruments statistics take the same values.

From this, we can compute wage gains for natives resulting from increased migration. For instance, if new migrants move to a city at a constant number of local residents and if this move increases the migrant share in total employment from the first quartile (decile respectively) of the distribution of the migrant variable across Chinese cities to the last quartile (decile respectively), then the associated natives' wage increase in the city is 10% (33% respectively). Around two-thirds of that figure comes from the complementarity of migrants with natives, and around a third from agglomeration effects induced by the increase in total employment generated by the migrant inflow. If migrants were to replace natives in local employment, thus keeping total employment density constant, the productivity of natives would increase by 6.4% (20.6% respectively) through the complementarity effect alone.

Finally, we test the theory prediction that the impact of migrants should be less positive, possibly negative, for those workers who are more substitutable by migrant labour in the production function. To do so, we split the native workers into skilled and unskilled workers, the categories being defined by each individual occupation, given that migrants are mostly unskilled. Skilled workers include senior management, professional and technical personnel, and office workers. First step estimation results for the two groups separately are displayed in Table 6 in Appendix C. Table 3 reports OLS and IV second-step estimations conducted on skilled and unskilled native workers separately. For the sake of comparison with the overall population of natives, we use the same control variables and instruments as in Table 2 column (8), our most complete specification. In keeping with the theory, we find that the migrant impact is less positive for unskilled than for skilled natives. Due to a lack of precision in the estimates, the difference is not fully significant from a statistical point of view, however. Still, when the overall impact of increasing the migrant share in total employment from the first quartile to the last quartile is 0.37 for skilled natives, it is only 0.27 for unskilled natives. The latter also experience slightly stronger agglomeration effects, at 0.14 instead of 0.11. Therefore the complementarity effect is 0.26 for skilled natives, while it is only 0.13 for unskilled natives. These findings confirm the intuition that skilled natives, who are less substitutable by migrants than unskilled natives, benefit more from migrants. Despite that, unskilled natives still gain from the presence of migrants, even if we only consider the complementarity effect.

---

the reasons detailed in section 1.1. If anything, an even larger overall impact of migrants is obtained.

4 Conclusion

This paper contributes to the literature on the impact of migration on local economies by investigating migration externalities in Chinese cities. China is an interesting case study because urbanisation has long been regulated by administrative means but labour mobility has sharply accelerated during the 2000s, feeding urbanisation and concomitantly raising concerns about the potential impact of migrant inflows on local residents' outcomes. Therefore, evaluating the role that migration plays in the process is a crucial step in assessing the possible scope for regional and urban policy in China.

Using 2007 microeconomic data from the National Bureau of Statistics, we find evidence of a large positive correlation of the city share of migrants with natives' wages. A number of instrumented estimations using different sets of instruments and control variables suggest that the effect is causal. The large total migrant impact (+10% when one moves from the first to the third quartile of the migrant variable distribution) can be attributed to gains from complementarity with natives in the production function (+6.4%, around two-third of the total), and to gains from agglomeration economies (+3.3%).[23] Overall, these results strongly support the nominal wage gains that can be expected from further migration and urbanisation in China.

Our findings on migration externalities provide evidence of rural migrants' contribution to China's economic growth and industrialisation. They support the hypothesis of a complementarity, rather than a crowding-out, that migrants bring to local workers. This is fully consistent with migrants, mostly from rural areas, being mainly concentrated in low-end labour-intensive industries that feed other local industries, thus contributing to an overall improvement of urban productivity. There are both institutional and population composition explanations for this, which have been documented in the literature (see e.g. Cai et al., 2008; Démurger et al., 2009; Deng and Li, 2010). First, institutional barriers are still high in the labour market and certain types of jobs are statutorily restricted to local residents. Second, because migrants are less educated and less skilled on average, they sort "naturally", making higher skilled local workers more productive by providing cheap labour to the low-skilled sector. Our estimations confirm this conjecture by providing some evidence of a larger positive impact for skilled natives, while the effect remains positive for the unskilled. Among unskilled workers, migrants are also little substitute to unskilled natives, probably because the precise occupations and tasks they hold remain different, although to a lesser extent than for skilled workers.

Beyond the overall large estimated impact of migrants on nominal wages for all types

[23]The two effects do not add up to the total overall effect because the model for wage levels is not linear.

of workers, their impact on occupation and location choice cannot be assessed with current Chinese data, and this remains to be investigated. While it probably played little role in the past, the feedback could become stronger in the future when migrants' education levels and mobility between sectors, in particular, are higher. Furthermore, whether migrants increase individual welfare is not under study here. To evaluate that, one should also estimate how the cost of living increases with the arrival of migrants and compare the nominal wage gains we exhibit with these costs. The literature on this topic is just starting to develop, even for Western countries, but it is clearly another hot issue to explore for China.

References

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

Bartik, T. (1991), *Who Benefits from State and Local Economic Development Policies?*, W.E. Upjohn Institute for Employment Research, Kalamazoo (MI).

Borjas, G. J. (2003), 'The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market', *Quarterly Journal of Economics* **118**(4), 1335–1374.

Borjas, G. J., Freeman, R. B. and Katz, L. F. (1997), 'How much do immigration and trade affect labor market outcomes?', *Brookings Papers on Economic Activity* **28**(1), 1–90.

Cai, F., Park, A. and Zhao, Y. (2008), The Chinese labor market in the reform era, *in* L. Brandt and T. Rawski, eds, 'China's Great Economic Transformation', Cambridge University Press, pp. 167–214.

Card, D. (2005), 'Is the new immigration really so bad?', *Economic Journal* **115**(507), F300–F323.

Chan, K. W. (2012), 'Migration and development in China: Trends, geography and current issues', *Migration and Development* **1**(2), 187–205.

Chen, Y., Démurger, S. and Fournier, M. (2005), 'Earnings differentials and ownership structure in Chinese enterprises', *Economic Development and Cultural Change* **53**(4), 933–958.

Ciccone, A. and Hall, R. (1996), 'Productivity, and the density of economic activity', *American Economic Review* **86**, 54–70.

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., Duranton, G. and Gobillon, L. (2012), 'The costs of agglomeration: Land prices in French cities', *CEPR Discussion Paper* **9240**.

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.

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.

Deng, Q. and Li, S. (2010), Wage structures and inequality among local and migrant workers in urban China, *in* X. Meng, C. Manning, S. Li and T. N. Effendi, eds, 'The Great Migration: Rural-Urban Migration in China and Indonesia', Edward Elgar, pp. 74–92.

Docquier, F., Ozden, C. and Peri, G. (2014), 'The wage effects of immigration and emigration in OECD countries', *Economic Journal* **124**(579), 1106–1145.

Duranton, G. and Puga, D. (2004), Micro-foundations of urban agglomeration economies, *in* V. Henderson and J.-F. Thisse, eds, 'Handbook of Regional and Urban Economics', Vol. 4, North-Holland, Amsterdam, pp. 2063–2117.

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.

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.

Jia, R. (2014), 'The legacies of forced freedom: China's treaty ports', *Review of Economics and Statistics* **96**(4), 596–608.

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.

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

National Bureau of Statistics (2008), *China City Statistical Yearbook*, China Statistics Press, Beijing.

Ottaviano, G. and Peri, G. (2012), 'Rethinking the effects of immigration on wages', *Journal of the European Economic Association* **10**(1), 152–197.

Peri, G. (2012), 'The effect of immigration on productivity: Evidence from U.S. states', *Review of Economics and Statistics* **94**(1), 348–358.

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.

Wu, W. and Gaubatz, P. (2013), *The Chinese city*, Routledge, London.

APPENDIX

A Summary statistics

Table 4: Summary statistics for local variables

	Mean	Std. Dev.	p10	p25	p50	p75	p90
Employment density (workers per sq. km)	414.2	426.6	25.44	61.76	315.1	665.0	938.1
log employment density	5.353	1.347	3.236	4.123	5.753	6.500	6.844
Area (sq. km)	2296.8	3098.0	537	835	1623	2718	4225
log area	7.337	0.867	6.286	6.727	7.392	7.908	8.349
Diversity	6.071	1.036	4.553	5.443	6.039	6.840	7.531
log diversity	1.789	0.175	1.516	1.694	1.798	1.923	2.019
Specialisation	0.170	0.0305	0.133	0.146	0.166	0.184	0.220
log specialisation	-1.789	0.175	-2.019	-1.923	-1.798	-1.694	-1.516
Market potential	162875	44600	99812	138752	167345	181594	215843
log market potential	11.96	0.288	11.51	11.84	12.03	12.11	12.28
Distance to seaport	521.5	414.6	82.75	179.2	428.0	721.4	1229.2
log distance	5.619	1.733	4.416	5.189	6.059	6.581	7.114
Migrant share	0.285	0.211	0.0842	0.140	0.208	0.361	0.624
log (1-migrant share)$^{-1}$	0.422	0.540	0.0879	0.150	0.233	0.448	0.978

Sources: National Bureau of Statistics (2008); 1% Population Census 2005.
Notes: 87 observations. Diversity is measured by the inverse of a sector concentration Herfindhal index, Specialisation is measured by the share in the local economy of employment in the firm's sector and Market potential is defined by the inverse-distance weighted sum of densities over all Chinese cities other than the city considered.

B First step variance analysis

Table 5 presents a full variance decomposition of the first step estimation reported in Table 6. It consists in computing, for each worker, the effect of a set of variables (by summing over the variables the estimated parameter times the value of the variable), and then the variance of this effect across all individuals and its correlation with the dependent variable. These computations allow measuring the explanatory power of each set of variables as well as its correlation with the other sets, so as to assess to what extent the observed effects are intertwined. Unsurprisingly, individual characteristics have the highest explanatory power. Their standard deviation (0.30) is half that of wages (0.75) and their correlation with wages (0.5) is the highest among all the effects (except residuals). Of particular interest here, city fixed effects also have a substantial explanatory power: The set of city dummies comes second after individual characteristics, with a similar standard deviation and a correlation with wages slightly lower. Sector dummies explain less than location, but ownership is fairly important. This fully corroborates the conclusions presented in section 3.1.

Table 5: Summary statistics for the variance decomposition

	Mean	St. dev.	Simple correlation with lwage
log wage	9.780	0.745	1
Effect of:			
Individual characteristics	1.477	0.303	0.497***
Sector dummies	0.0268	0.0969	0.243***
Enterprise ownership	-0.102	0.115	0.285***
City variables	8.378	0.315	0.398***
Among which:			
City fixed effects	8.490	0.314	0.398***
Specialisation	-0.112	0.0381	0.00441
Residuals	1.12e-10	0.555	0.744***
N	14,590		

Notes: The variance decomposition is based on estimation displayed in Table 6 column (1). *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

C First step estimation

Table 6: Individual wage disparities - OLS estimates for the first step

Log(wage)	All	Skilled	Unskilled
Male	0.252***	0.240***	0.293***
	(0.00992)	(0.0122)	(0.0165)
Years of education	0.0643***	0.0783***	0.0495***
	(0.00217)	(0.00275)	(0.00347)
Experience	0.0398***	0.0475***	0.0329***
	(0.00162)	(0.00208)	(0.00261)
Experience squared	-0.000838***	-0.000939***	-0.000773***
	(0.0000389)	(0.0000490)	(0.0000644)
Urban collective enterprises	-0.274***	-0.287***	-0.284***
	(0.0207)	(0.0271)	(0.0324)
Private or individual enterprises	-0.226***	-0.222***	-0.265***
	(0.0175)	(0.0289)	(0.0241)
Other ownership	-0.224***	-0.236***	-0.232***
	(0.0128)	(0.0163)	(0.0208)
Agriculture, mining	0.162***	0.133**	0.170***
	(0.0397)	(0.0563)	(0.0580)
Electricity, gas and water	0.238***	0.201***	0.297***
	(0.0328)	(0.0431)	(0.0526)
Construction	0.0975***	0.180***	-0.0236
	(0.0337)	(0.0432)	(0.0560)
Transport, storage, telecom	0.0926***	0.0706***	0.105***
	(0.0194)	(0.0268)	(0.0286)
Wholesale and retail trade	-0.0313	-0.0970***	-0.0594**
	(0.0207)	(0.0301)	(0.0245)
Finance and insurance	0.253***	0.202***	0.354***
	(0.0330)	(0.0394)	(0.0701)
Real estate	0.113***	0.132***	0.0260
	(0.0315)	(0.0405)	(0.0506)
Social services	-0.186***	-0.221***	-0.225***
	(0.0213)	(0.0295)	(0.0285)
Health, education, culture and research	0.0412**	0.0580***	-0.0163
	(0.0177)	(0.0212)	(0.0349)
Government and party agencies	0.0329*	-0.0210	0.0415
	(0.0186)	(0.0213)	(0.0502)
Specialization	0.0552***	0.0636***	0.0640***
	(0.0103)	(0.0140)	(0.0166)
Occupation dummies	Yes	No	No
City dummies	Yes	Yes	Yes
N	14,590	8,823	5,767
adj. R^2	0.442	0.399	0.350

Notes: Standard errors in brackets. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$. Experience is the actual number of years of work experience, i.e. participation to the labour force, reported in the survey. Reference groups: for occupation: other occupation (including soldiers); for enterprise ownership: state-owned enterprises; for economic sector: manufacturing.

D Feasible generalised least square (FGLS) estimator for the second stage

Table 7: The determinants of city effects - FGLS estimates for the second stage

	(1)	(2)	(3)	(4)
	FGLS1	FGLS2	FGLS3	FGLS4
Migrants	0.404***		0.367***	0.286***
	(0.041)		(0.042)	(0.046)
Density		0.102***	0.049**	0.072***
		(0.024)	(0.019)	(0.024)
Land area				0.103***
				(0.030)
Diversity				−0.077
				(0.131)
Market potential				0.154
				(0.098)
Distance to seaport				−0.010
				(0.013)

Notes: 83 observations for each regression. All regressors are logged. Standard errors in brackets. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

E Pairwise correlation coefficients for local variables and wage

Table 8: Pairwise correlation coefficients for local variables and wage

	Wage	Employment density	Area	Diversity	Market potential	Distance to seaport	Migrant share
Wage	1						
Employment density	0.37***	1					
Area	0.25*	-0.43***	1				
Diversity	0.22*	0.38***	0.08	1			
Market potential	0.43***	0.55***	-0.24*	0.06	1		
Distance to seaport	-0.33**	-0.39***	-0.04	-0.19	-0.38***	1	
Migrant share	0.71***	0.34**	0.18	0.26*	0.41***	-0.31**	1

Notes: All variables are in logarithm. The migrant share variable is defined consistently with the description given in section 1.1. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

F First stage regression for instrumented regressions

Table 9: First stage regressions for IV estimates, part I

	(1) IV1 Dens.	(2) IV1 Mig.	(3) IV2 Dens.	(4) IV2 Mig.	(5) IV3 Dens.	(6) IV3 Mig.
Land area			-0.525^{***}	0.108^{**}	-0.713^{***}	0.074
			(0.095)	(0.055)	(0.092)	(0.053)
Diversity			1.551^{***}	0.462^{*}	0.976^{**}	0.385
			(0.449)	(0.259)	(0.414)	(0.238)
Market potential			1.397^{***}	1.664^{***}	0.275	1.496^{***}
			(0.410)	(0.236)	(0.508)	(0.292)
Distance to seaport			-0.089^{*}	0.031	−0.036	−0.0071
			(0.052)	(0.030)	(0.045)	(0.026)
Peripherality	-1.104^{**}	0.204	0.620	1.777^{***}		
	(0.491)	(0.266)	(0.593)	(0.326)		
Henderson occupations	−2.982	5.077^{***}	3.419	4.614^{***}		
	(2.899)	(1.567)	(2.574)	(1.482)		
Doctors per capita	0.699^{**}	-0.379^{**}	0.714^{***}	0.070		
	(0.279)	(0.151)	(0.228)	(0.100)		
Telephones per capita	0.755^{***}	0.435^{***}			0.694^{***}	0.279^{***}
	(0.138)	(0.075)			(0.142)	(0.081)
Historic city					0.750^{***}	0.173
					(0.188)	(0.108)
Distance to historic city					−0.695	1.558^{***}
					(0.720)	(0.414)
Manufacturing share					0.347	0.291^{***}
					(0.240)	(0.138)
Non-agriculture share					−0.219	-0.786^{**}
					(0.255)	(0.147)
R^2	0.63	0.41	0.75	0.51	0.81	0.65

Notes: 83 observations for each regression. Standard errors in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Table 10: First stage regressions for IV estimates, part II

	(1) IV4 Dens.	(2) IV4 Mig.	(3) IV4 MP.	(4) IV5 Dens.	(5) IV5 Mig.	(6) IV5 Area
Land area	-0589^{***}	0.027	-0.049^{**}			
	(0.081)	(0.054)	(0.021)			
Diversity	1.229^{***}	0.288	-0.055			
	(0.416)	(0.277)	(0.106)			
Distance to seaport	-0.091^{**}	-0.034	-0.040^{***}			
	(0.045)	(0.030)	(0.011)			
Peripherality	-0.898^{**}	0.045	-0.983^{***}	3.144^{*}	3.301^{***}	-1.835
	(0.381)	(0.253)	(0.097)	(1.652)	(0.804)	(1.346)
Henderson occupations				0.465	5.859^{***}	7.549^{**}
				(3.578)	(1.741)	(2.916)
Doctors per capita	0.829^{***}	-0.073	0.196^{***}			
	(0.249)	(0.165)	(0.063)			
Telephones per capita	0.707^{***}	0.560^{***}	0.185^{***}			
	(0.124)	(0.083)	(0.032)			
Historic city				0.400^{*}	0.170	0.852^{***}
				(0.231)	(0.112)	(0.188)
Distance to historic city				-4.916^{***}	-3.427^{***}	3.125^{**}
				(1.853)	(0.902)	(1.510)
Manufacturing share				1.159^{***}	0.810^{***}	-0.178
				(0.337)	(0.164)	(0.275)
Non-agriculture share	-0.337	-0.720^{***}	-0.013	0.191	-0.766^{***}	0.036
	(0.233)	(0.155)	(0.059)	(0.371)	(0.180)	(0.302)
R^2	0.61	0.50	0.71	0.55	0.43	0.36

Notes: 83 observations for each regression. Standard errors in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.