Geography, Trade, and Internal Migration in China

Lin Ma Yang Tang *

June 29, 2017

Abstract

We quantitatively evaluate the income and welfare impacts of intercity migration in China. We develop a multi-city, multi-sector general equilibrium model with endogenous city and firm size distributions and imperfect labor mobility. We structurally estimate the model with data from 279 prefecture-level cities and real-world transportation networks. We find that intercity migration between 2000 and 2005 is able to explain 22 percent of the change in real income in the data. However, about 17.8 percent of the gain in real income is offset by the higher congestion costs in large cities. The gain from migration is concentrated in large cities with migration inflows, leaving more than half of the population with lower real income and the entire country with higher spatial inequality. We estimate the destination-specific entry barriers for some major Chinese cities, and find that the barriers often serve to protect the interests of the local residents at the expense of national welfare. We find that while internal trade liberalization reduces the spatial income inequality by discouraging migration, international trade liberalization induces more migration from inland to coastal cities and thus amplifies the spatial inequality. Migration in turn amplifies the gains from international trade by around 147 percent as it increases labor supply in coastal cities.

Keywords: regional trade; migration; welfare; economic geography

JEL Classification: F1; F4; R1; O4

*National University of Singapore (ecsml@nus.edu.sg) and Nanyang Technological University (tangyang@ntu.edu.sg), respectively. We thank Pol Antras, Davin Chor, Jonathan Eaton, Wen-Tai Hsu, Samuel Kortum, Andrei Levchenko, Michael Zheng Song, Kei-Mu Yi, Qinghua Zhang, Xiaodong Zhu, Thomas Zylkins, and the participants at ABFER Conference (2016), NBER China Group Meeting (May 2016), SMU Trade Workshop (2016), Asia-Pacific Trade Seminars (2016), Asia Meeting of the Econometric Society (2016), CUHK Workshop on Urban and Regional Growth in China (2016), East Asia Institute (2017), and University of Michigan (2017) for their helpful discussions and suggestions at various stages of this paper. We are solely responsible for the remaining errors.

1 Introduction

Over the past several decades, China has witnessed the largest wave of population migration in human history. After the easing of restrictions on household registration (Hukou) in the post-Mao era, an estimated 340 millions individuals (Chan [2011]) — roughly the entire population of the U.S. — have traveled thousands of miles from their hometowns in search of better working opportunities and potentially new lives. Undoubtedly, migration on this scale has altered the economic life of all Chinese citizens, migrants and non-migrants alike. Inflows of migrants bring fresh entrepreneurial ideas, an ample labor supply and increased consumption demand to the destination cities. At the same time, migrants compete with the natives for scarce resources and worsen the already mounting problems of congestion, pollution, and soaring housing prices in large cities. The controversial nature of the migration problem calls for comprehensive and quantitative assessments of the income and welfare impacts in each city due to migration. This paper attempts to offer some preliminary answers to these important questions.

We develop a multi-city, multi-sector quantitative framework with imperfect labor mobility to study the impacts of internal migration. Our framework is based on trade models with heterogeneous firms, following the ideas of Melitz [2003], Eaton et al. [2011], and di Giovanni and Levchenko [2012]. We extend this line of models with endogenous migration decisions. Individuals choose their preferred location depending on the relative real income and congestion dis-utility, geographic distance from their hometown, and idiosyncratic preferences. The city and firm size distributions, trade and migration patterns, and product and factor prices are all determined endogenously in the general equilibrium, enabling us to perform a series of rich quantitative studies. Inflows of migrants affect the destination cities through various channels. On the negative side, the native population suffers as migrants push down the nominal wage rates and increase congestion dis-utility. However, the increased labor supply also decreases the marginal costs of production for local firms and induces more firm entry, which reduces the ideal price index and benefits local residents. At the national level, population inflow into one city also benefits other cities through intercity trade: consumers and firms all over the country benefit from cheaper "imported" goods from the destination

cities subject to their relative locations on the transportation network. In the end, both the positive and negative effects of migration could prevail in our model, and thus, the welfare impacts of migration at both the city and the national levels are left to be investigated quantitatively.

We implement our model for 279 prefecture-level Chinese cities. We estimate the structural parameters with Simulated Methods of Moments (SMM) to match the data moments in input-output linkages, firm size distributions, internal and international trade patterns, and bilateral migration flows. To overcome the heavy computational load in the structural estimation, we develop a new iterative Particle Swarm Optimization (PSO) algorithm to efficiently estimate the model with large-scale parallel implementation. Our benchmark estimation successfully captures the key features in the data, such as the bilateral migration flows and city size distributions in both population and GDP.

We use the geographical locations of the cities — their relative positions on the road, railway, and waterway networks — to estimate the bilateral frictions in intercity trade and migration. Our estimation strategy follows Allen and Arkolakis [2014]. The estimation is based on high-resolution transportation maps and the transportation-mode-specific traffic volumes in each city. The resulting 279-by-279 geographic costs matrix is able to capture key features in the data on both intercity trade and bilateral migration patterns. We believe that the geographic costs matrix, and the associated migration costs matrix estimated in this paper can be widely applied to other China-related studies.

We study the impacts of intercity migration with counter-factual simulations. We find that the intercity migration between 2000 and 2005 increased the aggregate real income by 12.0 percent, which is about 22.2 percent of the real economic growth in the data. About 71.3 percent of the increase in the real wage is the direct result of migration: individuals moving from small and low-income cities into large and rich ones, to take advantage of higher real wages. The other 28.7 percent of the change is due to the agglomeration effects of migration. In both the data and our benchmark estimation, individuals tend to migrate toward large cities, leading to a more concentrated distribution of the population across space. This attracts more firms to enter large cities, which in turn offer more varieties to consumers at lower prices — a key mechanism shared by many models following Krugman

[1980]. The productivity gains in large cities also benefit the other cities through intercity trade along the transportation networks, leading to real income gains at the national level.

However, not all of the gains in real income can be translated into gains in welfare. The rapidly exploding population in large cities worsens the existing problems of congestion, such as higher housing prices, heavier pollution, and traffic jams. Naturally, these costs are borne by the people living in the large cities. The native population — those who choose not to migrate between 2000 and 2005 —have seen their real income offset by around 35.3 percent because of the soaring costs of congestion. At the aggregate level, around 17.8 percent of the gain in real income was offset by the change in congestion costs, and welfare only increased by around 9.9 percent between 2000 and 2005.

The gains in real income and welfare are not distributed evenly across cities, and internal migration leads to higher spatial inequality. The vast majority of the cities (around 50.6 percent of the total population) have become poorer due to migration: only 40 cities enjoyed higher real income and welfare, and they all received inward migration. This indicates that the positive effects of migration led by the reduction in the ideal price index quantitatively dominates the negative effects on nominal wage rates and congestion dis-utility in all 40 cities. Tombe and Zhu [2015] reach the opposite conclusion: in their model the regions that receive population inflow experience lower real income, and thus internal migration leads to lower spatial inequality. Our differences are mainly due to the endogenous firm entry and exit mechanism: the model in Tombe and Zhu [2015] abstracts from firm entry and productivity in each region becomes exogenous, which removes the benefits of agglomeration. In contrast, we allow for both positive and negative impacts of migration on real income in destination cities. Once we shut down the firm entry and exit mechanism in our model, we can qualitatively replicate the results in Tombe and Zhu [2015]. Our model predictions can be broadly supported in the data: after controlling for initial population and economic size, we find that cities with higher population inflow rates are associated with *higher* income growth between 2000 and 2005 in China.

The welfare gains in each city do not necessarily line up with the magnitude of the population inflows. Cities with relatively small inflows, such as Wuhan, Nanjing, Tianjin, and Suzhou, can still enjoy high welfare growth as they are strategically located either close

to large cities such as Beijing or Shanghai or at the crossroads of major traffic arteries. The location of these cities has made them the most direct beneficiaries of the productivity booms in the large cities via intercity trade: they are able to enjoy more varieties of tradable goods at lower prices.

We allow for asymmetries in bilateral migration frictions and use our model to estimate destination-specific migration barriers. The migration frictions in China are mainly due to the policy barriers to entry, in the form of the Hukou system, which can vary greatly across cities. For example, while migrants applying for Hukou in Beijing and Shanghai are usually required to have a college degree and pass certain income thresholds, these restrictions are virtually absent in smaller cities. These destination-specific barriers are also the focal point of policy debates. Assiduous city-planners and local residents often push to strengthen entry barriers to control explosive population growth in large cities. At the same time, the advocates of reform and migrants often decry the mere existence of these barriers as a violation of the basic human right of free mobility. We identify the destination-specific barriers from the discrepancies between the observed and the predicted migration flows without the barriers and estimate the entry barriers for China's four largest cities: Beijing, Shanghai, Guangzhou, and Shenzhen. We confirm that some larger cities are indeed harder to migrate to compared with the national average: the entry barrier into Shanghai is highest (22 percent higher than the national average), followed by Guangzhou (12 percent) and Beijing (6 percent). The barrier into Shenzhen is 1 percent *lower* than the national average, probably because of various policies that encourage immigration into the city. The high barriers in Beijing, Shanghai, and Guangzhou prevent further population inflow into these cities, resulting in a 0.74 to 3.57 percent loss in national welfare. However, the barriers do protect the native residents in the case of Shanghai: once removed, the population of Shanghai is predicted to increase by as much as 34.6 percent, resulting in a 6.87 percent *loss* in local welfare as the congestion disutility outweighs the gain in the real wage.

We also use our model to infer the destination-specific entry barriers and the associated city sizes that maximize either the local or the national welfare. We find that the city size that maximizes the local welfare is around 18 million for all four cities. This implies that as of 2005, the largest cities in China were underpopulated by between 6 and 47 percent because

of policy barriers, a result similar to the findings in Au and Henderson [2006]. The city size that maximizes the national welfare is usually around 22 million, which is significantly higher than the local optimum. This hints at a potential conflict of interest between the local and central governments, as the removal of these entry barriers is not a Pareto improvement. On the one hand, the central government prefers to lower the migration barriers in large cities, so that the aggregate welfare benefits from agglomeration effects, and spillovers via intercity trade. On the other hand, residents in large cities prefer a relatively high entry barrier, as they are the only people who bear the costs of over-population.

Lastly, we study how internal and international trade liberalization interacts with intercity migration. We find that lowering international trade barriers leads to 1) higher real income, 2) higher inequality in real income across cities, and 3) more migration across cities as compared to reductions in internal trade barriers of the same magnitude. A 10-percent reduction in internal trade frictions leads to a 4.5- percent increase in aggregate income and induces 0.5 percent of the entire population to migrate across cities. This is mainly because internal trade liberalization benefits small cities more than large ones, and thus reduces the between-city inequality in real income and mitigates the needs to migrate to large cities. In contrast, a 10-percent reduction in international trade barriers leads to a 20.0-percent increase in aggregate income and induces 7 percent of the entire population to migrate, mainly from inland cities to coastal ones with better access to foreign markets. This reallocation of workers across space amplifies the gains from trade by 147 percent as compared to a standard trade model without internal migration. Labor markets in the coastal cities are mainly responsible for the amplifications. Without migration, the higher labor demand in the coastal cities following liberalization quickly pushes up the local wage rate and the marginal costs of production, which restricts the extent to which firms can grow, and eventually limits the gains from trade. Once workers are allowed to move, the inflow of migrants pushes the labor supply curve outward, and thus decreases the equilibrium wage rates in the coastal cities. This allows the firms in these cities to grow larger, and the entire country to benefit more from international trade. Our results add to the recent debate on gains from trade following the work of Arkolakis et al. [2012]. We show that allowing for factor movements across space can amplify the gains from trade by a wide margin beyond what is often captured by the

overall openness. Firm entry/exit margin also amplifies the gains from international trade. Once we shut down both migration and firm entry channels, the gains in aggregate income from a 10-percent reduction in international trade barriers drop to 5.39 percent.

Our paper is most closely related to the quantitative works that focus on internal migration in China. Tombe and Zhu [2015] study how mis-allocation due to goods and labor market frictions affect aggregate productivity in China at the province level. Fan [2015] studies the impacts of international trade on skill premium at the prefecture-city level. While we share many modeling elements with their works, our paper highlights two important messages. Firstly, both of the previous studies are abstracted from firm entry and exit dynamics and thus the effects of agglomeration. We show that these elements indeed have substantial impacts on the aggregate gains from migration and trade liberalization. More importantly, allowing for firm entry reverts the negative relationship between population inflow and real wage in the destination regions as found in Tombe and Zhu [2015]. The positive relation between the two is also likely to be true in the data from our preliminary empirical analysis. Secondly, our work is the first to bring topography and real-world traffic networks into the study of migration. Our results highlight the importance of doing so: geographic locations of cities are, in many cases, central to their gains and losses during the massive migration.

This paper is also related to the growing body of literature that quantitatively examines the impacts of internal trade costs and migration costs separately or jointly on spatial distribution of economic activity within a country. Our theoretical framework is an extension of di Giovanni and Levchenko [2012] and di Giovanni et al. [2015], which present a multi-country and multi-sector model with heterogeneous firms and exogenous migration flows following Melitz [2003]. We apply their framework to a multi-city context and extend it by introducing an endogenous migration decision at the individual level. Caliendo et al. [2015] also recognizes the role of labor mobility frictions, goods mobility frictions, geographic factors, and input-output linkages in determining equilibrium allocations. They show that many quantitative results can be derived without the estimation of labor mobility frictions. Their result relies on the assumption that labor mobility frictions are constant over time, which is plausible in the case of the U.S. However, in the context of China, the reduction in migration frictions over time is widely believed to be the main driving force behind the

observed migration flows, and thus we need to model and estimate the frictions directly in our work.

Our work is broadly related to the large body of literature on the Chinese economy. Chow [1993] analyzes the path of development of different sectors in the economy. Brandt et al. [2008] further document the process of industrial transformation, the role played by institutions, and barriers to factor allocation. Hsieh and Klenow [2009] highlight how the mis-allocation of capital and output distortions have resulted in sizable losses in China's productivity. Song et al. [2011] argue that the reduction in the distortions associated with state-owned enterprises may be responsible for the rapid economic growth that began in 1992. Our work highlights the significance of internal migration in economic development. The reallocation of labor alone is able to explain a sizable proportion of the real income growth in China, which in turn implies the huge potential for economic growth of further easing the restrictions on labor mobility that still exist. Our analysis of optimal city size also reveals why reforms of migration policies are particularly difficult to implement: native residents in large cities lose welfare if the barriers are removed, and thus, such reforms might not be Pareto improvements. Our work also shows that a large proportion of income growth is offset by higher costs of congestion, especially in large cities: higher income does not necessarily translate into a higher degree of happiness in the case of China. This provides a new insight into the perception of China as an enigmatic country that simultaneously experiences spectacular economic growth, while being constantly bogged down by brewing social unrest and unhappiness.

The rest of the paper is organized as follows. Section 2 presents the theoretical model. Section 3 describes our quantification strategy. Section 4 discusses the main results. Section 5 concludes.

2 The Model

The production side of our model follows the multi-country trade framework in di Giovanni and Levchenko [2012]. We apply the model in a multi-city context and introduce an individual migration decision and labor market dynamics.

The economy contains a mass $\bar{L} > 0$ of individual workers, and $J > 1$ geographically segmented cities, indexed by $j = 1, 2...J$. The initial population distribution is given as $\{L_j^0\}$. Labor mobility across cities is allowed but is subject to frictions, which are specified later. There are two production sectors in each city j, namely, tradable and non-tradable sectors, which are denoted as sectors N and T, respectively. Individual workers obtain utility from the consumption of CES aggregate of intermediate goods produced in both sectors. Specifically, the utility function of an individual worker in city j takes the following form:

$$U_j = \left[\sum_{k \in \Omega_j^N} y\left(k\right)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\varepsilon\alpha}{\varepsilon-1}} \left[\sum_{k \in \Omega_j^T} y\left(k\right)^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\varepsilon(1-\alpha)}{\varepsilon-1}} - C(L_j), \quad 0 < \alpha < 1,$$

where ε represents the elasticity of substitution among all varieties and $y(k)$ is the quantity of variety k. Ω_j^s denotes the set of available varieties in city j and sector s. α captures the expenditure share on varieties produced in sector N_j. $C(L_j)$ represents the congestion dis-utility from living in city j, where L_j is the population size of city j. We assume that

$$C(L_j) = \rho \cdot L_j^{\phi},$$

and restrict $\rho > 0$ and $\phi > 0$ so that congestion dis-utility is increasing in city size.

The production of each intermediate good requires input bundles as inputs. To produce an input bundle in city j requires local labor and all of the available intermediate goods from sector N and T as inputs. The production technology for input bundles also varies between sectors to capture the idea that the relative contributions of labor and intermediate inputs in production may differ between sectors. Specifically, the production function for an input bundle in city j and sector s takes the form

$$b_j^s = L_{sj}^{\beta_s} \left[\left(\sum_{k \in \Omega_j^N} y\left(k\right)^{\frac{\varepsilon-1}{\varepsilon}} \right)^{\frac{\varepsilon\eta_s}{\varepsilon-1}} \left(\sum_{k \in \Omega_j^N} y\left(k\right)^{\frac{\varepsilon-1}{\varepsilon}} \right)^{\frac{\varepsilon(1-\eta_s)}{\varepsilon-1}} \right]^{1-\beta_s}, s = T \text{ or } N,$$

where b_j^s is the quantity of input bundles produced and L_{sj} is the employment in city j and sector s. In the tradable sector, the relative contributions of labor and intermediate goods

from sector N and T to production are β_T, $(1-\beta_T)\eta_T$ and $(1-\beta_T)(1-\eta_T)$, respectively. Similarly, in the non-tradable sector, the relative contributions of the three inputs are β_N, $(1-\beta_N)\eta_N$ and $(1-\beta_N)(1-\eta_N)$, respectively.

Given the specification of production technology for input bundles, it is straightforward to obtain the price of an input bundle in city j and sector s by solving the cost minimization problem:

$$c_j^s = w_j^{\beta_s}\left[\left(P_j^N\right)^{\eta_s}\left(P_j^T\right)^{1-\eta_s}\right]^{1-\beta_s}, \quad s = T \text{ or } N,$$

where P_j^N and P_j^T denote the ideal price indices in sectors N and T in city j, respectively. w_j is the wage rate in city j.

The intermediate goods market is featured in the fashion of monopolistic competition. Each intermediate good is produced by a single firm. Input bundles are the only input for the production of intermediate goods. Firms are heterogeneous in terms of their input bundle requirements for producing one unit of output. In other words, firms with higher productivity need fewer input bundles to produce one unit of output. Firms first need to pay f_e^s units of input bundles to enter sector s and city j. They then randomly draw their input bundle requirement a from a distribution function $G(a)$ from the following Pareto distribution:

$$G(\frac{1}{a}) = 1 - (a\mu)^\theta,$$

where $1/\mu$ denotes the maximum input requirement that a firm may draw. θ represents the tail index.

Once the productivity is realized, firms also need to choose which markets to serve. For a firm from city j to serve the market in city i, a fixed operating cost f_{ij} in terms of input bundles of city j must be paid.[1] Moreover, the standard iceberg trade cost assumption also applies to tradable intermediate goods here. To deliver one unit of intermediate goods from city j to city i, firms must ship $\tau_{ij} \geq 1$ units from city j.

[1]Firms in the non-tradable sector only need to decide whether to serve the local market or not. We can consider that f_{ij} to be infinity for firms in the non-tradable sector.

2.1 Firm's decision

We characterize the firms' optimization problem in detail in this subsection. Let X_i^s be the total expenditure in city i on goods produced in sector s. The standard CES utility function yields the following demand function for goods k and sector s from individual workers in city i

$$q_i^s(k) = \frac{X_i^s}{(P_i^s)^{1-\varepsilon}} \left(\frac{\varepsilon - 1}{\varepsilon} \frac{1}{\tau_{ij}} \frac{1}{c_j^s a(k)} \right)^{\varepsilon}.$$

A firm with input bundle requirement a in sector s and city j will serve city i if and only if the profit can cover the fixed operation cost, that is,

$$\pi_{ij}^s(a) \geq f_{ij} c_j^s, \tag{1}$$

where $\pi_{ij}^s(a)$ is the maximum profit level obtained from solving the following profit maximization problem:

$$\begin{aligned} \pi_{ij}^s(a) \equiv \max_{p_i^s(k)} p_i^s(k) q_i^s(k) &- a(k) \tau_{ij} q_i^s(k) c_j^s \\ s.t.\ q_i^s(k) &= \frac{X_i^s}{(P_i^s)^{1-\varepsilon}} \left(\frac{\varepsilon - 1}{\varepsilon} \frac{1}{\tau_{ij}} \frac{1}{c_j^s a(k)} \right)^{\varepsilon}. \end{aligned}$$

Standard results apply: the optimal pricing and the resulting sales revenue can be computed as follows:

$$\begin{aligned} p_i^s(k) &= \frac{\varepsilon}{\varepsilon - 1} \tau_{ij} c_j^s a(k), \\ R_{ij}^s(k) &= \frac{X_i^s}{P_i^{s1-\varepsilon}} \left(\frac{\varepsilon}{\varepsilon - 1} \tau_{ij} c_j^s a(k) \right)^{1-\varepsilon}. \end{aligned}$$

Moreover, by setting the inequality in equation (1) to be equal, we can derive the cutoff a_{ij}^s below which the firm in city j will serve city i:

$$a_{ij}^s = \frac{\varepsilon - 1}{\varepsilon} \frac{P_i^s}{\tau_{ij} c_j^s} \left(\frac{X_i^s}{\varepsilon c_j^s f_{ij}^s} \right)^{\frac{1}{\varepsilon - 1}}.$$

We assume that free entry holds in both sectors. The free entry condition in city j and sector s can be expressed as:

$$E\left[\sum_{i=1}^{J}\mathbf{1}\left(a\left(k\right)<a_{ij}^{s}\right)\left(\frac{X_i}{\varepsilon P_i^{1-\varepsilon}}\left(\frac{\varepsilon}{\varepsilon-1}\tau_{ij}c_j^s a\left(k\right)\right)^{1-\varepsilon}-c_j^s f_{ij}\right)\right]=f_e c_j^s.$$

Finally, the ideal price index in city i and sector s can be obtained as

$$\left(P_i^s\right)^{1-\varepsilon}=\sum_{j=1}^{J}\left(\frac{\varepsilon}{\varepsilon-1}\tau_{ij}c_j^s\right)^{1-\varepsilon}I_j^s\int_{-\infty}^{a_{ij}^s}a^{1-\varepsilon}dG\left(a\right),$$

where I_j^s denotes the firms entering sector s and city j.

2.2 Migration Decision

Labor mobility is allowed, subject to a certain migration cost. Each individual worker draws an idiosyncratic preference shock toward each city $\{\iota_i\}_{i=1}^{J}$, where ι_i is $i.i.d$ across locations and individuals. Therefore, the total utility from staying in city i includes two components: a common term shared all by individuals living in the city, and an idiosyncratic term that varies among individuals.

Migration across cities incurs some costs in terms of utility. In reality, when people migrate to a new city, they might suffer from homesickness or the adjustment to a new work environment. Let λ_{ij} denote the costs of migrating from city j to i. A worker living in city j will migrate to city i if and only if living in city i provides him with the highest utility among all J cities, that is,

$$U_i+\iota_i-\lambda_{ij}\geq U_k+\iota_k-\lambda_{kj},\forall k=1,2,...,J.$$

where U_i is the indirect utility from living in city i, which equals:

$$U_i=\left(\frac{\alpha w_i}{P_i^N}\right)^{\alpha}\left(\frac{(1-\alpha)w_i}{P_i^T}\right)^{(1-\alpha)}-C\left(L_i\right).$$

We assume that ι_i follows a Gumbel distribution with CDF:

$$F(\iota_i) = \exp\left(-\exp\left(-\frac{\iota_i}{\kappa}\right)\right),$$

where κ is the shape parameter. It is straightforward to show that conditional on U_i, the fraction of the population that migrates from city j to city i is

$$m_{ij} = \frac{\exp(\frac{U_i - U_j - \lambda_{ij}}{\kappa})}{\sum_{k=1}^{J} \exp(\frac{U_k - U_j - \lambda_{kj}}{\kappa})}.$$

The above equation is related to the "gravity equation" in international migration flows such as Grogger and Hanson [2011] and Ortega and Peri [2013]. Our functional form assumes that the bilateral migration flows are positively related to the per-capita income in the destination city, and negatively related to the bilateral frictions, which depend on the distance and policy barriers in our quantification in the next section. Both of these assumptions are strongly supported by the data in the context of international migration.

2.3 Equilibrium

Definition: Given a series of fixed costs, entry costs, trade costs, and migration costs $\{f_{ij}, f_e, \tau_{ij}, \lambda_{ij}\}$ in each city and sector, the equilibrium contains a series of prices $\{w_j, p_j^T(k), p_j^N(k)\}_{j=1}^J$, and a sequence of quantities $\{I_j^T, I_j^N, L_j, q_j^T(k), q_j^N(k)\}$ such that the following conditions hold:

(a) Individual workers maximize their utility by choosing locations and consumption bundles of goods from both sectors.

(b) Each intermediate goods producer maximizes its profits by choosing its price and quantity of output.

(c) The free entry condition holds in each city and sector.

(d) Goods market clearing:

$$\begin{aligned} X_i^N &= \alpha w_i L_i + (1-\beta_N)\,\eta_N X_i^N + (1-\beta_T)\,\eta_T X_i^T, \\ X_i^T &= (1-\alpha)\,w_i L_i + (1-\beta_N)\,(1-\eta_N)\,X_i^N + (1-\beta_T)\,(1-\eta_T)\,X_i^T. \end{aligned}$$

(e) Labor market clearing:

$$\sum_{j=1}^{J} L_j = \bar{L}.$$

3 Quantification of the Model

We quantify the model into 279 Chinese cities plus 1 location representing the rest of the world (ROW). All 280 locations can trade with each other. Individuals can migrate among the 279 Chinese cities subject to frictions, but migration between China and the ROW is not allowed. In the rest of this section, we first outline how we estimate the geographical structure, both within China and between China and the ROW, and we then describe the empirical issues in estimating the population distribution and bilateral migration flows in China. Lastly, we put the geographical structure and the population data together to calibrate and estimate the parameters of the model.

3.1 Estimating the Geographic Costs

3.1.1 Geography within China

As of 2005, there were 334 prefecture-level divisions in China. We focus on a selection of 279 prefecture-level cities in this paper because of data restrictions: our sample contains all of the cities that are included in both the *Chinese City Statistical Yearbooks* and the *One-Percent Population Survey* carried out in 2005 (thereafter 2005 Micro Survey). Our sample, which is illustrated in Figure 11, is representative: the 279 cities cover over 98 percent of the total population and over 99 percent of the total GDP in China in 2005. The vast majority of cities in China proper are included in our study; those missing are mainly the cities in Tibet, Xinjiang, and Inner Mongolia and various autonomous cities dominated by ethnic minorities

in southwest China.

We follow the approach in Allen and Arkolakis [2014] to estimate the matrix of geographic costs among the 279 cities, which is denoted as $\{T(i,j)\}$. Our estimation involves three steps. We first propose a discrete choice framework to evaluate the relative costs of trade using different transportation modes. Second, we discuss our approach to measuring the shortest distance between city pairs using different transportation modes. Third, we present our structural estimation strategy and discuss the estimated geographic costs matrix.

Suppose that there are M transportation modes indexed by $m = 1, 2...M$. For any pair of origin city j and destination city i, there exists a mass one of traders who will ship one unit of good. The traders choose a particular transportation mode to minimize the costs incurred from shipping. We assume that each trader k is subject to mode-specific idiosyncratic costs, which are denoted as ν_{km}. ν_{km} is $i.i.d$ across traders and transportation modes, and follows a Gumbel distribution $\Pr(e^{\nu} \leq x) = e^{-x^{-\theta_T}}$. The costs from j to i under mode m for trader k, $t_{km}(i,j)$, take the following form:

$$t_{km}(i,j) = \exp(\psi_m d_m(i,j) + f_m + \nu_{km}), \tag{2}$$

where $d_m(i,j)$ is the distance from city j to i using transportation mode m. ψ_m is the mode-specific variable cost, f_m is the mode-specific fixed cost, and ν_{km} is the trader-mode specific idiosyncratic cost. The specifications above allow us to explicitly identify the fraction of traders from city j to i using transportation mode m, which is identical to the fraction of trade flows under mode m:

$$\frac{\exp(-a_m d_m(i,j) - b_m)}{\sum_{n=1}^{M}(\exp(-a_n d_n(i,j) - b_n))}, \tag{3}$$

where $a_m = \theta_T \psi_m$ and $b_m = \theta_T f_m$. We next estimate the mode-specific distance matrix $d_m(i,j)$. We start with the high-resolution transportation maps from the *2005 China Maps* published by Sino Map Press. Each raster image has 4431-by-4371 resolution, so each pixel roughly corresponds to a $1.3km$-by-$1.3km$ square. We then assign a cost value to every pixel on the map to indicate the relative difficulty of traveling through the area using a specific

transportation mode. For example, on the map to measure normalized road network costs, we assign pixels with no road access a cost of 10, pixels with highways a cost of 2.5, pixels with national-level roads a cost of 3.75, and pixels with provincial and other types of road access to be 6.0. All of the costs are chosen to roughly reflect the differences in speed limits under Chinese law.[2] As in Allen and Arkolakis [2014], we normalize the pixels with navigable waterways, including open seas, with a cost of 1, and all other pixels with a cost of 10. To construct the raster for normalized railroad cost, we assign all pixels with rail road access a cost of 1, and all-other a cost of 10. We then identify the central location of each of the 279 cities on the raster maps, and apply the Fast Marching Method (FMM) algorithm between all pairs of cities i and j to obtain a normalized distance between them for each transportation mode, $d_m(i, j)$.

Given the mode-specific distance matrix, we next estimate the cost parameters $\{a_m, b_m\}$ in Equation 3. Following Allen and Arkolakis [2014], we estimate these parameters by matching the fraction of trade volume in each city and the transportation mode in the data. We construct the city-mode-specific trade volume from two data sources. From *China City Statistics Yearbook 2005*, we are able to observe the quantity shipped in metric tons in each city using transportation mode m. For instance, the total quantity shipped in city i by railroad includes goods shipped from city i to all other cities and the those delivered to city i from all other cities by railroad. Next, we turn to the transaction-level custom dataset for China to estimate the relative value per ton of goods under different modes of transportation. In 2005, the results from 22.82 million custom transactions indicate that the goods shipped via railroad and sea command low values at only 408 and 489 RMB per ton, respectively. The goods shipped via road are valued much higher at around 2,450 RMB per ton. Combining the quantity and value information, we can construct the fraction of trade volume under each transportation mode in all cities.

In the model, the total trade volume of city i by transportation mode m, denoted as

[2]On average, the speed limit on highways is 120 KM/H, that on national-level roads is 80 KM/H, and that on provincial-level roads 50 KM/H.

$V_m(i)$, equals

$$V_m(i) = \sum_{j=1}^{J} \exp(-a_m d_m(i,j) - b_m) + \sum_{j=1}^{J} \exp(-a_m d_m(j,i) - b_m).$$

The share of total trade volume using transportation mode m in city i, $s_m(i)$, can thus be expressed as

$$s_m(i) = \frac{V_m(i)}{\sum_{n=1}^{M} V_n(i)}. \tag{4}$$

We estimate $\{a_m, b_m\}$ using a non-linear least square routine to minimize the distance between the simulated $\{s_m(i)\}_{i=1}^{J}$ and the data counterpart. We search over 100,000 initial points for $\{a_m, b_m\}$ in our algorithm to avoid local minimum. In the end, our estimated $\{a_m, b_m\}$ is able to capture the main feature of the data, as presented in Table 1. In the data, the vast majority of intercity trade is carried out via road transportation (76.3 percent), and the same applies in our model (75.4 percent). We are also able to capture the relative weight of rail and river transportation with error margins at around 1 percentage point.

	Model	Data
Average share by road	0.754	0.763
Average share by rail	0.152	0.155
Average share by river	0.094	0.083

Table 1: Model Fit in Estimating Geographic Costs

Note: The table presents the average share of trade volumes via different modes across all of the cities. The model results are based on equation (4). The data counterparts are computed from the Chinese City Statistics Yearbooks and the Custom Dataset.

For the value of θ_T, we follow the estimations in Allen and Arkolakis [2014] and set it to 17.65.[3] Given $\{a_m, b_m\}$ and θ_T, the discrete choice framework implies that the average

[3]The estimation of θ_T requires bilateral trade flow data, which do not exist in the case of China. However, directly using the value estimated from the U.S. data is almost innocuous. Equation (5) shows that θ_T serves two purposes. Firstly, it scales $T(i,j)$. When we use $T(i,j)$ to estimate the bilateral trade and migration costs in the next section, we use the Chinese data to discipline the scale of the matrix, and thus directly adopting θ_T is harmless. Secondly, θ_T also serves as the elasticity of substitution between different modes of transportation, as both a_m and b_m are linear functions of θ_T. The elasticity is inherent to the transportation technology, and thus is unlikely to vary across countries.

geographic costs from city j to i can be obtained as follows:

$$T(i,j) = \frac{1}{\theta_T}\Gamma\left(\frac{1}{\theta_T}\right)\left(\sum_m \exp(-a_m d_m(i,j) - b_m)\right)^{-\frac{1}{\theta_T}}, \tag{5}$$

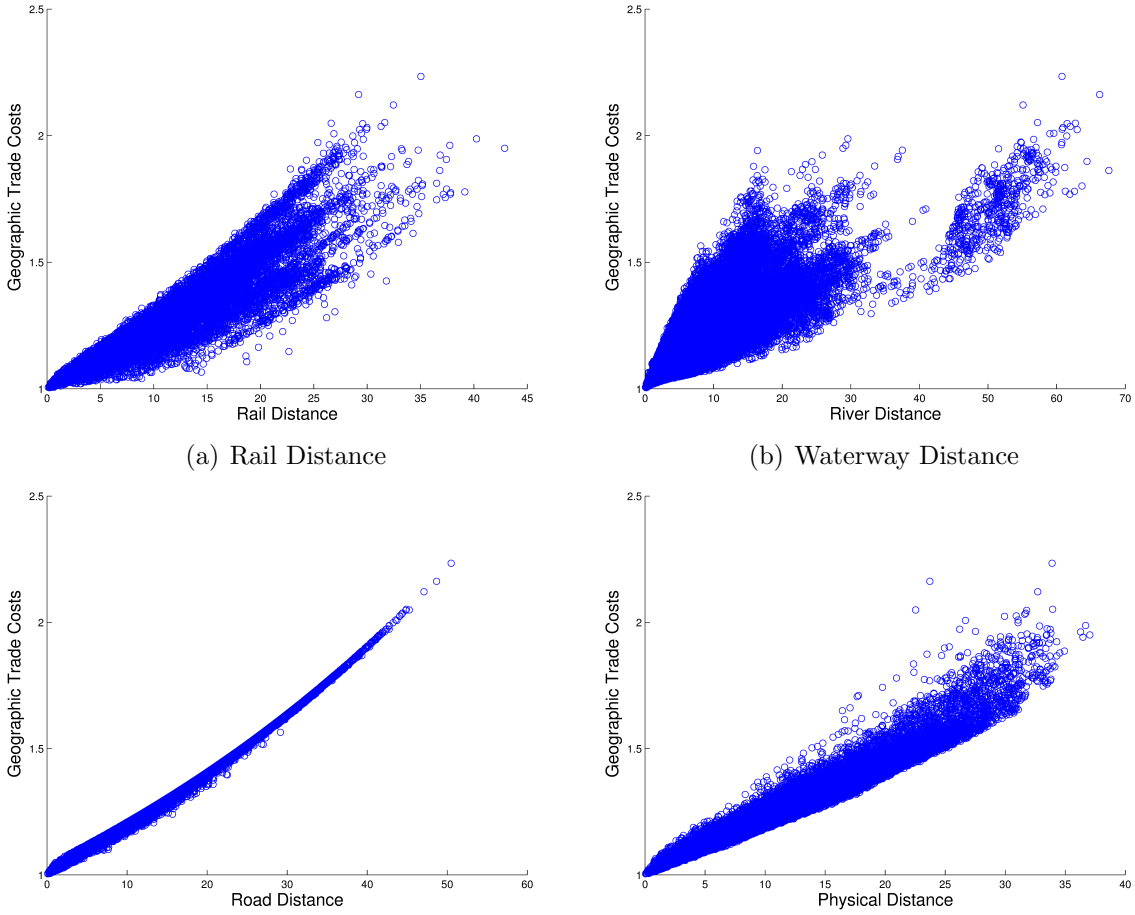
where $\Gamma(\cdot)$ is the standard Gamma function.

We plot the estimated geographic costs matrix, $T(i,j)$, against different measures of distance, $d_m(i,j)$, in Figure 1. Unsurprisingly, the trade costs increase with distance regardless of the transportation mode. Of the three modes of transportation, the estimated T matrix mostly depends on the length of the road network because all of the cities in our sample have access to the national road system, and the vast majority of intra-China trade goes through the road network. In contrast, geographical costs can vary significantly between city pairs with similar rail or waterway distances. Traveling by river or coastal sea has the largest variation, mainly because a large proportion of Chinese cities do not have easy access to any waterway.

Empirical works often use physical distances between cities as proxies for transportation costs, implicitly assuming that city pairs with similar physical distance also share similar difficulties in transportation. We plot our T matrix against the physical distance in the last panel of Figure 1. Unsurprisingly, the geographic costs increase with physical distance. However, conditional on a given physical distance, the variations in geographic costs are large and increasing with physical distance. For example, the geographical costs for city pairs 1,000km apart could range between 1.15 and 1.37; for pairs 3,000km apart, the variations can range between 1.54 and 2.0. This indicates that physical distance is at best a noisy proxy for the costs of transportation. However, the extent to which using geographical distances can refine the existing empirical findings remains an open question to be explored by future research.

3.1.2 Geography between China and the World

We condense the 148 trading partners of China into the location of the rest of the world (ROW). The choice of trading partners is again, because of data restrictions: all of the countries included in the *World Development Index* (WDI), *COMTRADE*, and our sea



(a) Rail Distance

(b) Waterway Distance

(c) Road Distance

(d) Physical Distance

Figure 1: Geographic Trade Costs by Transportation Mode

Note: The four panels above plot the estimated geographical costs matrix, T, against the mode-specific measures of distance obtained by FMM. The last panel plots the T matrix against the physical distance between two cities. The physical distance is measured as the great circle distance between city centers. The physical distances are normalized such that the distance between Beijing and Tianjin (110.9 KM) is 1.

distance database (which we discuss later) are included in the sample. We estimate the geographical distances following a similar strategy with a few modifications.

First, we assume that the ROW and China can only trade through water transportation. This assumption is again because of data restrictions: while shipping route data between major ports in the world are widely available, much less can be obtained for the other two modes of transportation. This is also an innocuous assumption: records from Chinese customs indicate that on average, over 80 percent of international trade measured in value

and over 90 percent measured in weight is shipped by sea.[4]

We then measure the waterborne distance between the ROW and every coastal city in China. We start by collecting shipping route data from `www.sea-distances.org`. For each country k, we pick its largest port and then measure the shortest shipping distance between this port and a given coastal city i in China, which is denoted as r_{ik}.[5]. The distance between ROW and the coastal city i is then computed as

$$d_{\text{sea}}(i, \text{ROW}) = \xi \cdot \left[\sum_{k=1}^{148} \left(\frac{\Lambda_k}{\sum_{j=1}^{148} \Lambda_j} \right) \cdot r_{ik} \right].$$

ξ converts nautical miles, which is the unit of s_{ik}, to the units used in $d_{sea}(\cdot)$ for waterborne transportation in China.[6] The terms in the square brackets are the average shipping distance between all of the ROW ports and the coastal city i weighted by the trade volumes between country k and China. Lastly, we use Equation (5) again with the $d_{\text{sea}}(i, \text{ROW})$, assuming the distances in the other two modes to be equal to infinity, to compute the T_{ij} between any coastal city in China to the ROW.

For inland city j in China, we first measure its distance to the nearest coastal city, $i(j)$, with the estimated T matrix above and assume that the inland city will trade with the ROW through the nearest coastal city. Therefore, the geographic distance between any inland city j and the ROW is $T_{i(j),j} \cdot T_{\text{ROW},i(j)}$, where $T_{i(j),j}$ is the distance between inland city j and its nearest coastal city and $T_{\text{row},i(j)}$ is the distance between the coastal city $i(j)$ and the ROW. See the appendix for more details on extending the geography to include the ROW.

3.2 Population and Migration

We use the population distribution over the 279 Chinese cities in the year 2000 as the initial population distribution in our benchmark model. We back-out the structural parameters

[4]The authors' own calculation using custom data from China between 2000 and 2005.

[5]For countries facing multiple oceans or with long coast-lines, such as the U.S., Canada, and Russia, we pick multiple ports facing different directions and take the average. The shortest shipping distance is the minimum distance across different routes: direct shipping, going through the Suez Canal, the Panama Canal, the Strait of Gibraltar, etc.

[6]We compare the distance in nautical miles between Guangzhou, Shanghai, and Dalian to the respective distances in $d_{sea}(\cdot)$ matrix computed above. We then define ξ as the average across the three ratios.

on migration costs and congestion disutility from the bilateral migration flows between 2000 and 2005. The estimation depends on two sources of data: 1) the population census in 2000, which provided detailed population counts at the county (sub-prefecture) level, and 2) the 2005 micro survey, which recorded the current location and the location in 2000 for each respondent. Conceptually, it is straightforward to construct both the initial population distribution in 2000 and the bilateral migration matrix between the two years using the information above. However, directly using these data will lead to problematic estimates.

The main challenge is that the official definitions and boundaries of cities changed constantly between 2000 and 2005. This means that cities in the two data sources, even those with exactly the same name, are not directly comparable. The 279 cities we use are based on the 2005 definition. Out of this sample, 49 cities did not exist as prefecture-level administrations in 2000, and 12 cities changed their boundaries significantly. To solve these problems, we construct a geographically-consistent dataset of city populations between 2000 and 2005 based on the city boundary defined in 2005 ("2005-cities" hereafter). The official records from the central and provincial governments contain information on how sub-city administrative units (counties) are grouped into new cities or how they are re-assigned among existing cities. We use these records to map counties in 2000 to their respective cities in 2005 and then reconstruct the populations of 2005-cities based on this county-city mapping. The resulting data set is the first geographically-consistent population panel data at the city level.

We use the total population of the 148 trading partners of China as the raw population of the ROW.[7] We allow for potential differences in total factor productivity (TFP) between the ROW and China by introducing a parameter to measure the relative efficiency between

[7]The population data source is *World Development Indicators,2000.*

Chinese and ROW workers. The initial population used in the benchmark simulation is

$$L_{2000} = \begin{bmatrix} \ell_1 \\ \ell_2 \\ \vdots \\ \ell_{279} \\ A \cdot \ell_{\text{ROW}} \end{bmatrix},$$

where $\ell_i, i = 1, \cdots 279$ are the populations of the Chinese cities in 2000 that we constructed above, ℓ_{ROW} is the total population of the 148 trading partners, and A is the relative TFP that we estimate later.

3.3 Quantifying the Structural Parameters

Our parameter space contains the following structural parameters:

$$\{\epsilon, \theta, \mu, \beta_N, \beta_T, \eta_N, \eta_T, \alpha, \kappa, f_e, A\},$$

and three origin-destination-specific matrices $\{f_{ij}, \lambda_{ij}, \tau_{ij}\}$. We calibrate some of the parameters based on the common approaches in the literature, and structurally estimate the rest.

3.3.1 Calibration

ϵ is the elasticity of substitution among all of the intermediate goods in the final goods production. This parameter generally ranges from 3 to 10 in the literature, and we pick the middle value of 6. θ is the tail index of the firms' productivity distribution. In our model, the firms' employment follows a power law distribution with a tail index of $\theta/(\epsilon - 1)$. We follow di Giovanni and Levchenko [2012] by setting θ to be 5.3 so that the tail is equal to 1.06, the value documented in Axtell [2001]. The values of β_N and β_T reflect the share of labor in total output, and we calibrate them using *China 2002 Input-Output Table*. We use the basic flow tables of 42 industries and compute $\beta_N = 0.47$ and $\beta_T = 0.33$ as the ratios between the total wage bills and the total output in the non-tradable and tradable sectors,

respectively. η_N and η_T are the share of non-tradable intermediate goods in non-tradable and tradable sectors, and we also calibrate them using China 2002 input-output table. The data suggest that $\eta_N = 0.42$ and $\eta_T = 0.22$. Similar to what di Giovanni and Levchenko [2012] documented using U.S. data, intermediate goods from non-tradable sectors play a larger role in the production of other non-tradable goods in the Chinese data as well. In contrast to the U.S. data, non-tradable goods are overall less important in both sectors, probably because many services industries, such as finance and consulting, are relatively less developed in China. α governs the expenditure share on non-tradable goods. We set it to be 0.61, the share of total consumption of non-tradable goods, which is computed from the final use table in the input-output table from the same year.

For the fixed operating costs matrix f_{ij}, we first turn to the 2005 micro survey, and approximate $1/f_{ii}$ by using the fraction of entrepreneurs in each city among all working population. Following di Giovanni and Levchenko [2012] we set the off-diagonal elements, f_{ij}, as the sum of the two diagonal elements f_{ii} and f_{jj}. At this stage f_{ij} matrix is not yet in the unit of local labor, and to convert it to the correct unit, we again follow di Giovanni and Levchenko [2012] by scaling the entire matrix with a factor ζ. We set ζ to ensure interior solutions in all of the counter-factual simulations. We summarize all of the calibrated parameters and their corresponding targets in Table 2.[8]

Para.	Targets	Para. Value
β_N	labor share in non-tradable sectors	0.47
β_T	labor share in tradable sectors	0.33
η_N	non-tradable share in non-tradable sectors	0.42
η_T	non-tradable share in tradable sectors	0.22
α	expenditure share on non-tradable goods	0.61
θ	Pareto index in emp. distribution	5.3
ϵ	elasticity of substitution	6.0

Table 2: Calibrated Parameters

Note: The calibration targets for β_s, η_s, and α come from the 2002 Chinese input-output table for 42 industries. The target for θ comes from Axtell [2001] and the values for ϵ and ζ come from di Giovanni and Levchenko [2012].

[8]Interior solution here means $a_{ij} <= 1/\mu$, where $1/\mu$ is the theoretical upper bound of the unit cost distribution. We calibrate ζ such that the number of entering firms is about twice the size of the number of operating firms in the benchmark model to guarantee that not all firms that enter choose to operate.

3.3.2 Estimation

We jointly estimate the other elements of the parameter space, $\{\tau_{ij}, \lambda_{ij}, \kappa, f_e, \rho, \phi, A\}$, with structural estimation. We first reduce the dimension of the space by reducing the two matrices, τ_{ij} and λ_{ij}, to a few parameters, and then estimate these parameters with SMM following the ideas in McFadden [1989] and McFadden and Ruud [1994].

We first simplify the τ_{ij} matrix with the geographic costs matrix estimated from the previous section, T. We assume that the iceberg trade costs take the following form:

$$\tau_{ij} = \begin{cases} \bar{\tau} \cdot T_{ij} & , \text{ if } i \neq \text{ROW and } j \neq \text{ROW} \\ \tau_{\text{row}} \cdot \bar{\tau} \cdot T_{ij} & , \text{ if } i = \text{ROW or } j = \text{ROW} \\ 1 & , \text{ if } i = j \end{cases}$$

The first line assumes that the iceberg trade costs between Chinese cities are proportional to the geographic costs matrix. As widely documented in the trade literature, national borders usually introduce significant costs to international trade.[9] We allow for an additional international trade barrier, τ_{row}, to capture the border effect, and we later use it to carry out policy experiments. The above simplifications reduce the estimation of the entire τ_{ij} matrix down to the estimation of two scalars: $\bar{\tau}$ and τ_{row}.

We model the migration costs matrix λ_{ij} as:

$$\lambda_{ij} = \begin{cases} \left(\bar{\lambda} \cdot T_{ij}\right) \cdot \delta_i & , i \neq j \\ 0 & , i = j \end{cases}$$

The migration costs are affected by two parts. The first part, $\bar{\lambda} \cdot T_{ij}$, is symmetric between i and j and proportional to the geographic trade cost T_{ij}. All else being equal, it is generally easier to move to nearby cities because of ease of travel and similarities in language, cuisine, and climate. The literature estimating the "gravity equation" of international migration, such as Grogger and Hanson [2011] and Ortega and Peri [2013], also found that the physical distance significantly reduces the migration flow, and thus shall be considered as part of the frictions to migration. Our geographic cost matrix (T) is estimated using the traffic

[9] See McCallum [1995] and Anderson and van Wincoop [2003] for examples.

volumes of goods instead of passengers, and we have omitted air transportation all together. However, directly using the T matrix is largely innocuous for two reasons. First, the relative importance of the road, railway, and waterborne transportation for passengers is roughly the same as for goods.[10] Second, air transportation for passengers is negligible, and only constitutes less than 0.8 percent of total traffic between 2000 and 2005, according to *China City Statistical Yearbooks*.

The second part is the *destination-specific* migration cost, δ_i. A large part of migration costs in China comes from the policy barriers from entry in the form of the Hukou system, which often vary greatly across cities. For example, while migrants applying for Hukou in Beijing and Shanghai are usually required to have a college degree and pass certain income thresholds, these restrictions are absent in smaller cities. By introducing destination-specific barriers, our model is able to quantify the relative difficulties of moving to certain cities and later to use δ_i to carry out counter-factual policy experiments. For computational reasons, we cannot separately estimate δ_i for each of our 279 cities. Instead, we focus on "tier-1" cities with populations higher than 10 million: Beijing, Shanghai, Guangzhou, and Shenzhen.

The vector of the 12 parameters to be estimated by SMM is summarized as

$$\Theta = \{\bar{\tau}, \tau_{\text{ROW}}, \bar{\lambda}, \kappa, f_e, \rho, \phi, \delta_{\text{Beijing}}, \delta_{\text{Shanghai}}, \delta_{\text{Guangzhou}}, \delta_{\text{Shenzhen}}, A\}.$$

Our estimation strategy is to find the vector $\widehat{\Theta}$ such that

$$\widehat{\Theta} = \text{argmin}_{\Theta}[S - \widehat{S}(\Theta)]\widehat{W}[S - \widehat{S}(\Theta)]'. \tag{6}$$

S is a vector of data moments that we explain in detail later in this section, $\widehat{S}(\Theta)$ is the counter-part moments generated by the model, which depends on the input parameter Θ, and $\widehat{W}$ is the weighting matrix.[11] The model is computationally heavy to evaluate, and therefore we use an iterative particle swarm optimization (PSO) algorithm to take advantage of large-scale parallel computing power in solving the minimization problem. We provide the details

[10] In 2005, 91.5 percent of passenger transportation goes by road, followed by 6.7 percent by railroad and 1 percent by rivers. For goods transportation, the ranking is the same (see Table 1). Data source for passenger traffic is the same as goods traffic: *China City Statistical Yearbooks*.

[11] In the benchmark model we use the identity matrix as weighting matrix.

of our algorithm in the appendix.

The S vector contains 20 data moments that are important in disciplining the Θ vector. The first element in the S vector is the average intercity-trade-to-GDP ratio. We estimate the overall volume of intercity trade using the Investment Climate Survey in China from the World Bank [2005]. This survey covers 12,500 firms in 31 provinces of China, and it asks the firms to report the percentage of their sales by destination: within the city limit, within the province, within China, and overseas. On average, 62.5 percent of the total sales of the firms surveyed are generated outside of their home city, and thus we use this as the internal-trade-to-GDP ratio. This data moment helps to identify $\bar{\tau}$, the magnitude of the iceberg trade costs matrix.

The second moment is the average number of firms in the largest 20 cities by population. We estimate this to be around 84,400 based on the Second Economic Census carried out in 2004.[12] This statistics helps to identify f_e, which captures the fixed costs of entry in input bundles. We assume that this parameter is the same across cities. Inherently, it captures the cost paid to reveal one's ability as an entrepreneur, which is unlikely to be affected by the differences in infrastructures and institutions across cities. In equilibrium, the costs of input bundles differ across cities, and thus the de facto costs of entry in the unit of the numeraire actually vary across cities.

The next two moments focus on international trade. We target 1) the trade openness of China in 2005, defined as (exports+imports)/GDP, and 2) the relative size of China and the ROW in 2005. The first moment can be obtained from the National Bureau of Statistics, and is around 59.4 percent. The second moment is based on the data from WDI, which states that the 148 trading partners combined are around 21.32 times larger than China in terms of GDP. These two moments help identify the international trade barrier, τ_{row}, and the relative productivity, A.

The other moments in the S vector are based on the bilateral migration flows between 2000 and 2005 that we have previously estimated. We denote the migration matrix M in

[12]We interpret "legal entities" as firms.

the data as follows:

$$M_{2000,2005} = \begin{bmatrix} \ell_{11} & \ell_{12} & \cdots & \ell_{1J} \\ \ell_{21} & \ell_{22} & \cdots & \ell_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ \ell_{J1} & \cdots & \cdots & \ell_{JJ} \end{bmatrix}, \tag{7}$$

where ℓ_{ij} indicates the population flow from city j to city i between 2000 and 2005, and $J = 279$ is the number of Chinese cities. Note that the summation of each column of the matrix gives the population distribution across cities in 2000, while the summation of each row produces the population distribution in 2005. We denote the population distribution vector in these two years as L_{2000} and L_{2005}, respectively. We focus on the following groups of moments based on this matrix.

We first target the overall magnitude of intercity migration as captured by the aggregate stay-rate, which is the proportion of the population that choose not to move:

$$\text{Aggregate Stay Rate} = \frac{\sum_{i=1}^{J} \ell_{ii}}{\sum_{i=1}^{J} L_{2000}(i)}.$$

Stay-rates can also be computed for each city separately to measure the propensities to emigrate out of the city:

$$\text{Stay Rate}_i = \frac{\ell_{ii}}{L_{2000}(i)}.$$

Stay-rates differ significantly across cities in the data. For example, fewer than 0.01 percent of the individuals living in Beijing in 2000 migrated to other cities in 2005, but the emigration rate for some smaller cities was as high as 49 percent. To capture this feature, we rank cities by their population in 2000 and group cities into four categories: top 10, top 20, top 40, and all of the others. We then target the average stay-rates within each group of cities separately. In a similar vein of logic, we also target the standard deviation of city-specific stay-rates, both across the entire nation and within each of the four groups of cities.

We use inflow rates to capture the propensities to move into a certain city i:

$$\text{Inflow Rate}_i = \frac{\sum_{j \neq i}^{J} \ell_{ij}}{L_{2005}(i)}.$$

This indicates the percentage of the population that moved into the city i after 2000 as of 2005. We use the inflow rate information along several dimensions. At the aggregate level, we target the correlation between the logarithm of city population in 2000 and the city-specific inflow rate. This is mainly to capture the feature in the data that cities with higher initial populations usually have higher inflow rates as well. We also target the inflow rates of the four cities on which we have imposed destination-specific entry barriers, δ_i: Beijing, Shanghai, Guangzhou, and Shenzhen. In addition to the identification of δ_i, the inflow rates of the largest cities also help us to identify the parameters governing congestion disutility, which is not directly observable in the data. Our identification comes from the assumption that congestion only becomes a severe discomfort in the large cities, and it can thus be inferred from the changes in the populations of these cities. We also target the Pareto tail index of the city size distribution in 2005 to capture the overall shape of the population distribution across space. All the data moments are summarized in Table 7. All of the estimated parameters, along with the standard errors, are reported in Table 3. The standard errors are estimated with 200 repetitions of bootstrapping. See the appendix for the details of the bootstrapping algorithm.

3.4 Model Fit

We evaluate the fitness of our quantification on both the targeted and the un-targeted moments. Table 7 reports the 20 targeted moments, in the data and in our benchmark quantification. Overall, our model is able to match the data moments with relatively small discrepancies. For most of the moments, the relative differences between the model and the data are smaller than 5 percent. On average, the difference between the model and the data moments is 10 percent.

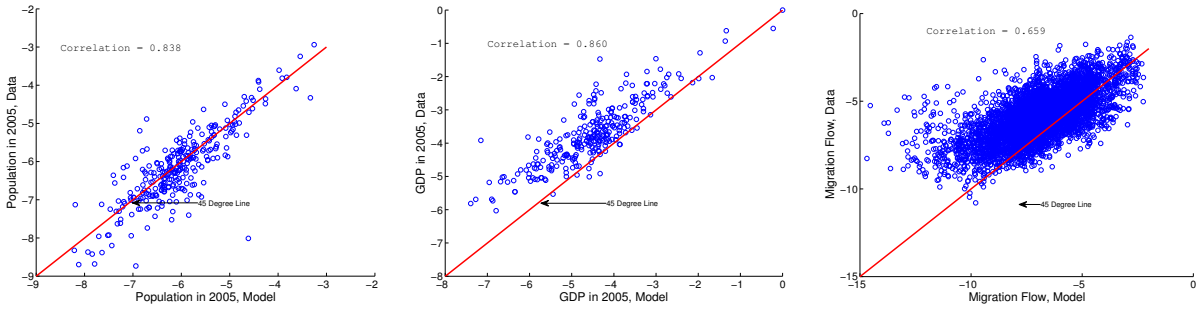
We also check the fitness of our model by examining the un-targeted moments in the data: the population and GDP distributions in 2005 and the entire bilateral migration flow

Para.	Value	S.E
$\kappa \times 1000$	2.140	0.016
f_e	2.780	0.087
$\bar{\lambda} \times 1000$	14.373	0.121
$\bar{\tau}$	2.135	0.070
ϕ	3.472	0.017
ρ	193.977	2.117
τ_{row}	0.098	0.003
A	0.121	0.001
$\delta_{Beijing}$	1.060	0.006
$\delta_{Shanghai}$	1.225	0.007
$\delta_{Guangzhou}$	1.122	0.005
$\delta_{Shenzhen}$	0.986	0.006

Table 3: Parameters, Estimated

Note: This table reports the results of the estimation using SMM. The standard errors are computed with 200-repetition bootstraps. κ is the parameter that governs the distribution of idiosyncratic location preferences; f_e is the fixed cost of entry; $\bar{\lambda}$ is the scale of the migration frictions; $\bar{\tau}$ is the scale of the iceberg trade costs; ϕ and ρ are the parameters governing the congestion disutility; τ_{row} is the international trade barrier; A is the relative TFP between China and the ROW; and $\delta_{\cdot}$ is the city-specific migration barrier.

matrix between 2000 and 2005. These results are summarized in Figure 2.



(a) 2005 Population (b) 2005 GDP (c) 2000-05 Migration Flow

Figure 2: Model Fit: Un-targeted Moments

Note: The graphs above plot the population, GDP distribution, and bilateral migration flows implied by the model against their counter-parts in the data. In all of the graphs, the total population of China is normalized to be 1, and we plot the logarithms of the population and migration flows. The data source for both graphs is the 2005 mini survey.

Panel (a) plots the population distribution for 2005 as predicted by the model against its data counterpart. They are consistently lined up along the 45-degree line with a correlation coefficient of 0.838. Similarly, we are also able to match the city-level GDP with a correlation

of 0.86, as shown in Panel (b). In Panel (c), we scatter plot the bilateral migration flow between all city-pairs with positive flows in the data. Our calibration strategy performs reasonably well: the correlation between the model and the data is 0.659. The magnitude of bilateral migration flows are mostly influenced by the specification of the λ matrix, which we assume to be roughly proportional to the geographical trade costs. The model fit validates our assumption: when individuals choose migration destinations, neighboring cities and cities located along main traffic networks serve as better targets because of easier access. The model tends to slightly under-predict when the migration flow in the data is relatively small. Given that we only target several moment conditions for the migration matrix, we view these discrepancies as an affordable price to pay for our parsimonious quantification strategy.

4 Quantitative Results

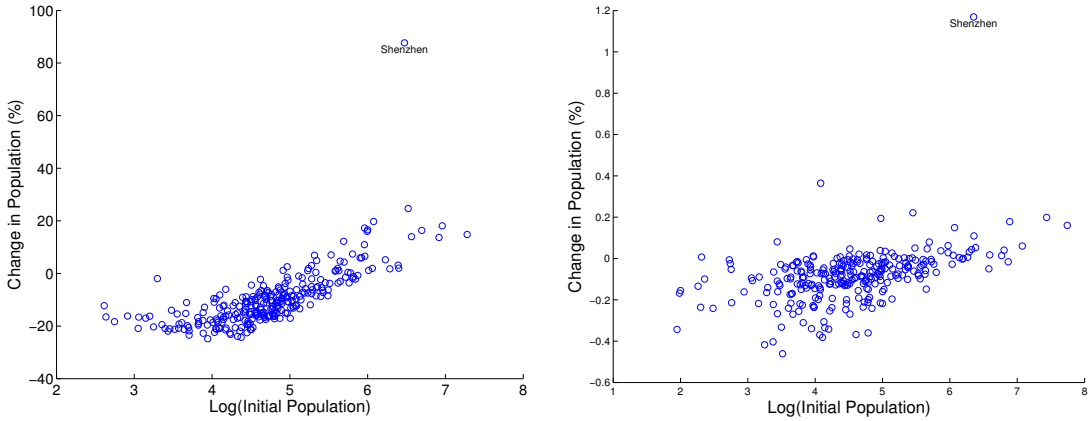
In this section, we first evaluate the impacts of migration on the equilibrium outcome, and then study how internal migration interacts with internal and international trade liberalization. Lastly, we perform a series of policy experiments in which we lower the destination specific barriers, and we use these experiments to shed light on the optimal city size.

4.1 The Impacts of Migration

We simulate a counter-factual situation by setting the migration cost multiplier $\bar{\lambda}$ to a sufficiently large value so that individuals will not migrate. This effectively sends all of the migrants in our benchmark model back to their 2000 locations. We compare the results to our benchmark quantification to study the impacts of intercity migration between 2000 and 2005.

We first study the pattern of intercity migration. As shown clearly in Figure 3, cities with larger initial population in 2000 also tended to attract higher population inflows between 2000 and 2005, both in the data and in the model. The population of the largest cities, such as Beijing and Shanghai grew by around 15 percent. The population of Shenzhen—the city built almost from scratch since the beginning of the reform— effectively doubled its size

during the period. The population inflow rates in the model are similar to those in the data. In contrast, smaller cities generally lost a proportion of their population to large cities: in the data, 233 out of the 279 cities experienced emigration, and in our model, 239 cities. The concentration of the population in the largest cities is the most important feature of the data that drives most of the results in our counter-factual analysis in this section, as we describe in detail. Interestingly, the pattern of migration found here – from smaller and arguably less productive locations to larger and more productive ones – is also reflected in cross-border migrations as documented in di Giovanni et al. [2015].



(a) Model (b) Data

Figure 3: Population Change and Initial Population

Note: The figure plots the change in population between 2000 and 2005 against the initial population in 2000. Panel (a) is our benchmark simulation, and Panel (b) is the data.

The results are summarized in Figure 4. At the aggregate level, the migration pattern described above increases the national real wage by about 12.01 percent. Over the same period, the real GDP of China increased by around 54 percent in the data.[13] This implies that around 22.2 percent of the economic growth, a surprisingly large number considering that the aggregate productivity is kept constant between the counter-factual and the benchmark economy, can be explained by the re-allocation of population across the country. The gain in real wage comes from several channels. Firstly, the marginal product of labor is usually higher in larger cities, and thus, moving workers into those cities leads directly to real economic growth. Second, concentrations of the population in large cities increase consumption

[13]Source: Penn World Table 8.1.

demand and lower the marginal costs of production by reducing the costs of labor in local markets. This leads to more firm entry in large cities, increasing the number of varieties available and lowering the ideal price index. Third, the economic boom in the large cities can also benefit all of the other cities through intercity trade: consumers in the other cities benefit from the decreased prices of the goods produced in the large cities, and firms benefit from the lowered costs of intermediate goods. The first channel mainly benefits the migrants themselves, while the second and third channels benefit both the migrants and those who stayed in their home cities.

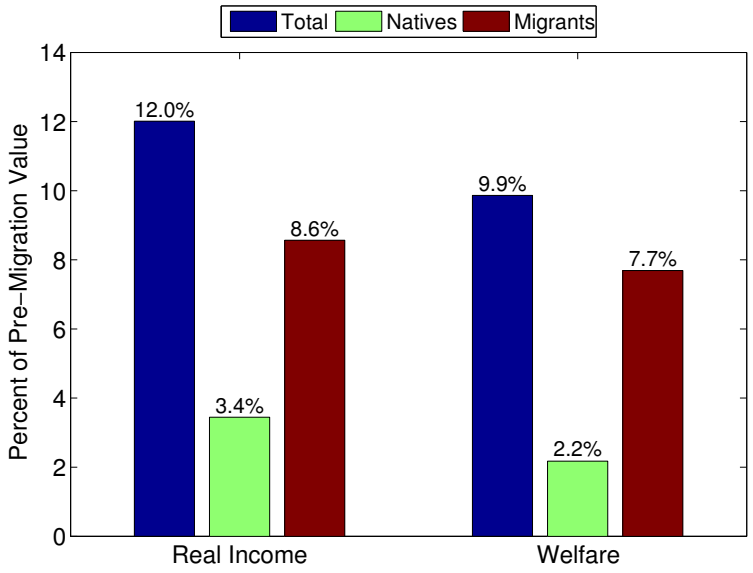


Figure 4: Aggregate Impacts of Migration

Notes The graph plots the changes in total real income and total welfare between 2000 (pre-migration) and 2005 (post-migration). The values are presented as percentage changes to the pre-migration value. The difference between real income and welfare is the congestion dis-utility.

To shed light on the relative importance of the potential channels, we decompose the gains in real income by two subgroups of individuals: the migrants and the stayers. The two groups of individuals can be affected through migration differently. Intuitively, migrants benefit most directly from a higher real wage — after all, the differences in real wage are the main reason for their migration in the first place. The native population can both benefit and suffer from inflows of migrants: population inflow lowers the ideal price level, which benefits the natives; however, migrants also tend to dampen the nominal wage in the destination cities, which affects the natives negatively. To evaluate the impacts, we first note that the

change in aggregate real income, ΔY, can be written as

$$\Delta Y = \sum_{i=1}^{J} \left[\left(\frac{w_{i,2005}}{p_{i,2005}} \right) \cdot \sum_{j=1}^{J} \ell_{ij} \right] - \sum_{i=1}^{J} \left[\left(\frac{w_{i,2000}}{p_{i,2000}} \right) \cdot \sum_{j=1}^{J} \ell_{ji} \right],$$

where ℓ_{ij} is the population flow from city j to city i between the two years. ΔY can be rearranged as

$$\Delta Y = \sum_{i=1}^{J} \ell_{ii} \cdot \left[\frac{w_{i,2005}}{p_{i,2005}} - \frac{w_{i,2000}}{p_{i,2000}} \right] + \sum_{i=1}^{J} \sum_{j \neq i} \ell_{ij} \cdot \left[\frac{w_{i,2005}}{p_{i,2005}} - \frac{w_{j,2000}}{p_{j,2000}} \right].$$

The first part of the above decomposition is the change in real wage for the stayers, and the second part is that for the migrants. This decomposition suggests that around 28.7 percent of the gain in real income can be attributed to the stayers, resulting in a 3.4 percent aggregate increase in the real wage for them. The other 71.3 percent is contributed by the migrants, resulting in an 8.6 percent increase in total real income for the migrants. Note that migrants only accounts for around 17.3 percent of the entire population, which implies that in per-capita terms, the gain in real wage for a typical migrants is $8.6 * 82.7/(3.4 * 17.2) = 12.16$ times more than that for a typical stayer.

The downside of higher population concentrations in large cities is higher congestion dis-utility. While the real wage has increased by 12.0 percent, the total welfare in China only increased by around 9.86 percent between 2000 and 2005, implying that around $1 - 9.86/12 \approx 17.8$ percent of economic growth is offset by higher congestion. The burden of higher congestion dis-utility is mainly borne by those living in the large cities. Similar to the decomposition that we performed for the case of real income, we decompose the change in total welfare in the following:

$$\Delta U = \sum_{i=1}^{J} \ell_{ii} \cdot \left[u_{i,2005} - u_{i,2000} \right] + \sum_{i=1}^{J} \sum_{j \neq i} \ell_{ij} \cdot \left[u_{i,2005} - u_{j,2000} \right].$$

Similar to the previous case, the first part summarizes the change in total utility for the stayers, and the second part for the migrants. After taking the congestion dis-utility into account, the native population sees its welfare increase by only 2.2 percent. Compared to

the 3.4 percent increase in the real wage, around $1.2/3.4 \approx 35.3$ percent of the economic benefit was offset by higher congestion dis-utility. Migrants suffer from higher congestion dis-utility to a lesser degree. The welfare of the migrants increased by 7.7 percent, which is $1 - 7.7/8.6 \approx 10.5$ percent lower than the change in real income.

Intercity trade only marginally amplifies the impacts of migration. We perform another set of counter-factual analysis in which inter-city trade is not allowed.[14] Once internal trade is shut down, the same magnitude of reduction in $\bar{\lambda}$ improves aggregate income by around 10.7 percent, which is only $1 - 10.7/12.01 \approx 11$ percent lower than the benchmark case in which intercity trade is allowed. The weak link between internal trade and migration is mainly due to the substitutability between the mobility of goods and population in our model. Internal migration allows the consumers of tradable goods to live closer to the site of production, depressing the need for intercity trade. This implies that even if intercity trade is completely excluded, the gains from migration will only be marginally affected. Similarly in our benchmark exercise when we reduce the migration frictions, the overall trade openness *declined* between 2000 and 2005 in our model from 72.69 percent to 68.12 percent, again supporting the substitutability between trade and migration. Similar pattern will again show up in section 4.3, and we will discuss this issue in more details there.

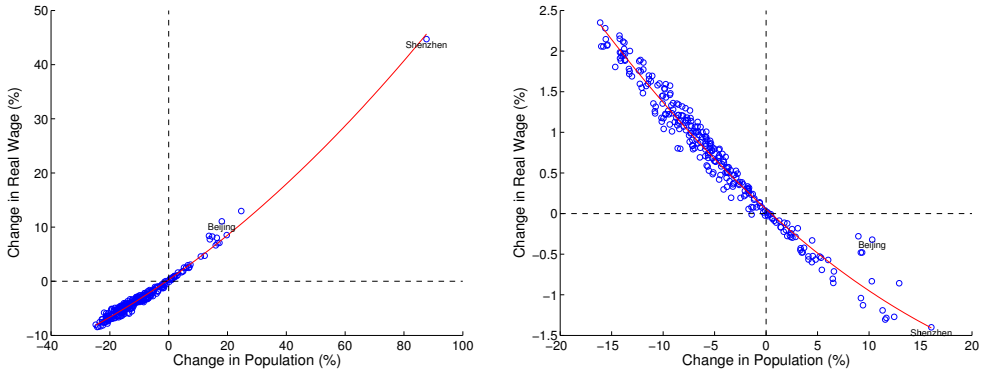
City-level Impacts The aggregate gains in real wage are not uniformly distributed. Of the 279 cities in our sample, only 40 cities experienced real wage growth, while the other 239 cities suffered loss. This implies that those who stayed in the 239 cities between the years, which is around 50.6 percent of the entire population, might have suffered a lower real wage because of migration. The first panel in Figure 6 shows that the 40 "winners" of migration are mostly the large and coastal cities in the eastern part of the country. As shown in the third panel of Figure 6 , they are also the only 40 cities that received positive population inflow over the period. This implies that, contrary to the results found in Tombe and Zhu [2015], we find that real income in regions receiving population *increases*, and thus internal migration actually leads to *higher* spatial inequality.

[14] We first set both $\bar{\lambda}$ and $\bar{\tau}$ to some sufficiently high value such that no migration will take place and internal trade is shut down. We then restore $\bar{\lambda}$ to its value in the benchmark model and study the aggregate impacts of migration in an autarky economy.

The reason that we reach different conclusions from Tombe and Zhu [2015] is because we adopt different quantitative frameworks. We build on a Krugman model which allows for both positive and negative impacts of migrants on local real income. Inflows of migrants tend to dampen the nominal wage, which negatively affects the local residents. However, population inflow also induces more firm entry, which lowers the price index and thus benefits the locals instead — essentially the "love of variety" mechanism from Krugman models. Our quantitative results show that the extensive margin of firm entry dominates the negative impacts on nominal wage, and thus the positive relationship between population flows and real income prevails. On the other hand, the model in Tombe and Zhu [2015] follows and Eaton and Kortum [2002] framework, which does not allow for firm entry. Moreover, the measure of technology, (T_i in both Eaton and Kortum [2002] and Tombe and Zhu [2015]'s terminology) is exogenously fixed, which implies that the benefit of agglomeration and population inflow is absent. As the relationship between inflow of migrants and the welfare of local residents is the focal point of interest in policy debates on migration, both internal and international, we deem that the question requires a more careful investigation.

To highlight the difference between our work and Tombe and Zhu [2015], we simulate another counter-factual in which the number of firms in each city and sector is fixed to its value in year 2000 and then repeat the benchmark exercises by decreasing $\bar{\lambda}$. At the aggregate level, the same reduction in $\bar{\lambda}$ only leads to a 2.82 percent aggregate real income gain once we shut down firm entry and exit. This implies that around $1 - 2.82/12.01 \approx 77$ percent of the gain in real income is due to the extensive margin of firm entry and exit. More importantly, once the extensive margin of firm entry and exit is shut down, inflows of migrants tends to lower real income in destination cities, and thus lower overall spatial inequality. Figure 5 plots the change in real wage against the change in population in simulations with and without firm entry. When firm entry is allowed, the "love of variety" dominates the negative impacts on nominal wages, and cities with higher inflows enjoy larger gains in the real wage. Once the entry channel is excluded, the relationship is reversed: inflows of migrants *decrease* the real income of the destination cities — a finding similar to Tombe and Zhu [2015].

Our quantitative results suggest that population inflow tends to induce firm entry, and thus improves living standards in the destination cities. We empirically test both channels



(a) With Firm Entry and Exit (b) Without Firm Entry and Exit

Figure 5: Population Flows and Change in the Real Wage

Notes: The graph plots the change in real income against the change in population for two cases. The first panel is our benchmark case, in which firm entry and exit is allowed. In the second panel, we shut down firm entry and exit.

in the following through some simple OLS regressions.

Table 4 reports the regression results, in which we have the net population inflow rate, initial population, and initial economic size at the city level in year 2000 as regressors, and the percentage changes in the number of firms and per capita GDP between 2000 and 2005 as dependent variables.[15] The only possible data source to compute "number of firms" in each city across the years is the *Annual Surveys of Manufacturing Firms*, and we use these surveys here as well. A well-known limitation of the data is that only state-owned firms and private firms whose sales revenue exceeds a certain threshold are surveyed each year, and thus the "number of firms" in the estimation loosely refers to the "number of large firms" instead. Nevertheless, the truncation issue is mitigated as 1) we measure the changes over time, instead of the levels themselves, and 2) in each city the total number of firms is highly correlated with the the number of large firms in the survey in the census year 2004, in which we can observe the total number of firms at the city level through economic census. The correlation coefficient reaches a high level at 0.88. With or without the controls for initial conditions, higher inflows rates are always associated with higher growth rates in the number of firms and per capita GDP, and the elasticity are sizable at around 0.574 to 0.689. Our

[15]The sample size (227) is smaller than our quantitative sample of 279 due to the changes in the definition of prefecture-level cities between 2000 and 2005. Unlike in the population data where we can observe units at sub-city level and reconstruct the distribution, we do not have the detailed data for the number of firms. For more details, see the discussion in Section 3.2.

	$\Delta \log(\text{GDP/POP})_{2000,2005}$			$\Delta \log(\text{N.Firms})_{2000,2005}$		
	(1)	(2)	(3)	(4)	(5)	(6)
Net Inflow Rate	0.312***	0.261**	0.574***	0.879***	0.766***	0.689***
	(0.110)	(0.108)	(0.161)	(0.129)	(0.136)	(0.179)
Initial Population		0.044*	0.176***		0.097**	0.065
		(0.026)	(0.060)		(0.038)	(0.065)
Initial GDP			-0.133***			0.033
			(0.044)			(0.053)
Constant	0.476***	0.257**	1.478***	0.470***	-0.015	-0.314
	(0.022)	(0.125)	(0.377)	(0.029)	(0.195)	(0.534)
N	227	227	227	227	227	227
R-squared	0.028	0.036	0.103	0.111	0.136	0.134

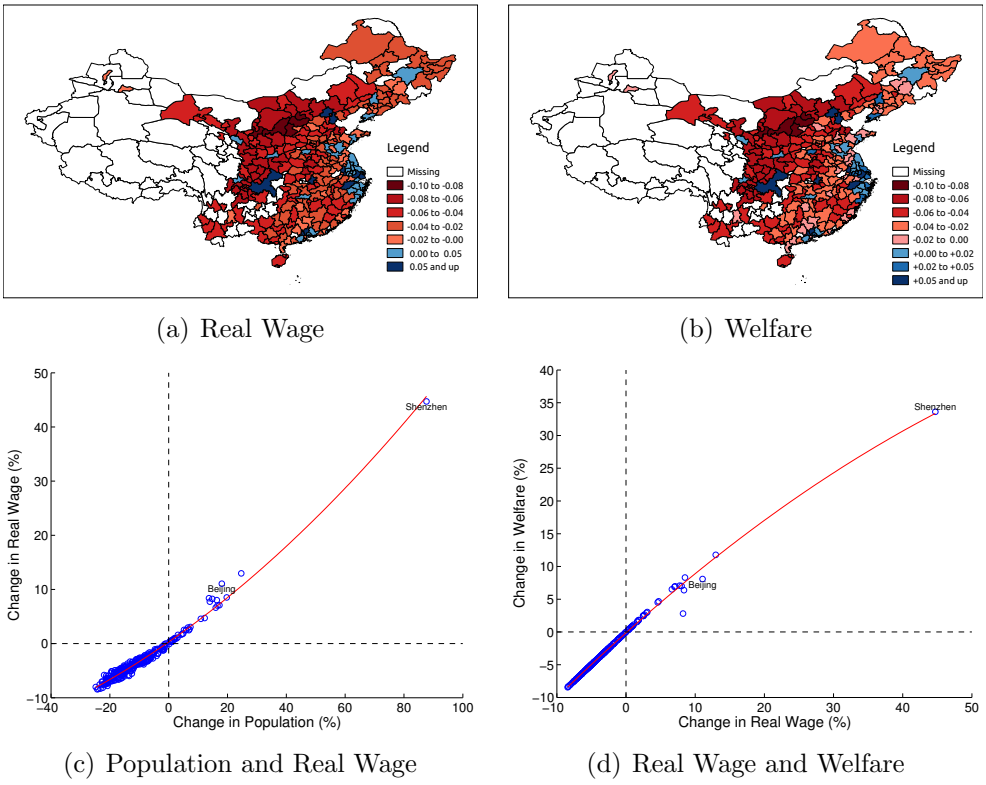
Robust standard errors in parentheses.
* p<0.10, ** p<0.05, *** p<0.01

Table 4: Net Inflow Rate, Changes in Per Capita GDP and Number of Firms

Data source: *China City Year Books* and *Annual Surveys of Manufacturing Firms*. Initial population and GDP refers to the data in year 2000. Net inflow rate is the net changes in population divided by the initial population.

simple empirical test does not claim causality between population inflow, firm entry, and income growth, and we leave a full-fledged empirical investigation on the causality issues to future research. Nevertheless these results suggest that the negative relationship between the two variables obtained in Tombe and Zhu [2015] and our counter-factual economy without firm entry/exit is unlikely to hold true in the data. Using geography and culture distance as instruments for openness to immigration, Ortega and Peri [2014] found that immigration leads to higher per-capita income in the context of international migration, which is similar to our finding in the case of China.

Similar to the case at the national level, not all of the gains in real income can be translated into higher welfare. Large cities pay the price of immigration in terms of higher congestion dis-utility, leaving the overall gain in welfare much smaller. The most striking case is Shanghai, the largest city in our sample. Between 2000 and 2005, the population of Shanghai increased from 14.5 to 16.8 million. The real wage in Shanghai increased by about 9.4 percent as a result. However, the population growth imposes a congestion dis-utility equal to around 6.6 percent of the pre-migration real wage, wiping out $6.6/9.4 \approx 70.2$ percent



(a) Real Wage

(b) Welfare

(c) Population and Real Wage

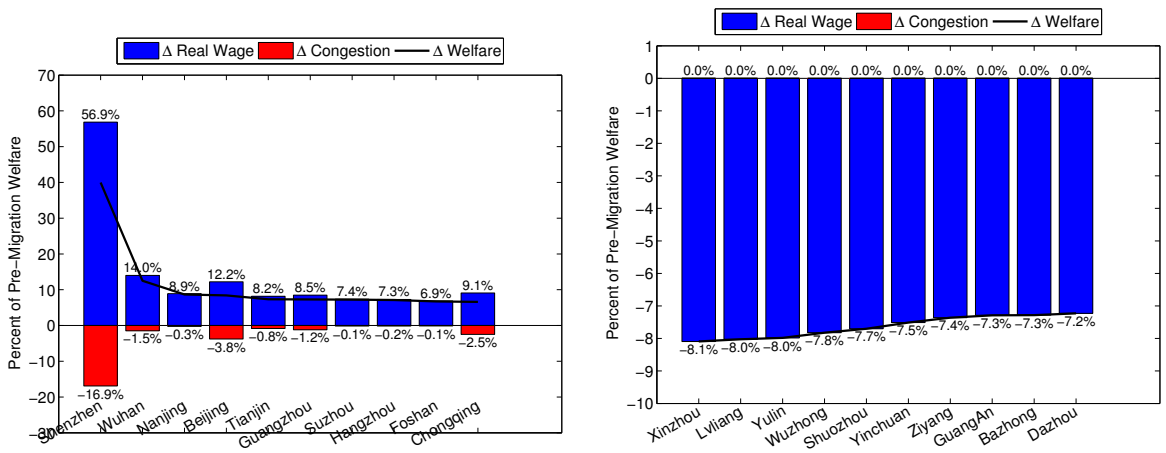
(d) Real Wage and Welfare

Figure 6: Real Wage and Welfare

Note: The graphs above plot the change in real income and welfare between 2000 and 2005 due to inter-city migration implied by the model. The difference between real income and welfare is the congestion dis-utility. For more details, refer to Section 4.

of economic gain. Similarly, higher congestion dis-utility offsets the real income growth in Beijing by 31.1 percent, and in Shenzhen by 29.7 percent. The impacts of congestion are more muted in smaller cities. For example, the changes in congestion roughly equal 10.7 percent of the real income gain in Wuhan, and 1.35 percent in the case of Suzhou. Nevertheless, despite the surge in congestion costs, all of the 40 cities that received net population inflows enjoyed higher welfare, hinting that there are still potential gains in the local welfare of the destination cities from migration friction reductions, a topic that we discuss in detail later in the section. All of the the cities that experienced population outflow suffer lower welfare, meaning that despite the 9.9 percent gain in aggregate welfare, a large fraction of the population still suffers from lower welfare because of migration outflows.

Figure 7 plots the top 10 winners and losers in terms of welfare. All of the top 10 winners are migration inflow-cities, and the top winner is Shenzhen, with a net welfare gain of 40



(a) Top 10 Winners

(b) Top 10 Losers

Figure 7: Winners and Losers from Migration

Note: The two panels above plot the changes in the real wage and congestion costs as a percentage of welfare before and after migration. The first panel plots the top 10 cities in terms of percentage change in welfare, and the second panel plots the bottom 10 cities according to the same measure.

percent, followed by Wuhan, Nanjing, and Beijing. Shenzhen experienced the largest inflow of migrants, both in absolute numbers and percentage terms. Between 2000 and 2005, the population of the city increased by 140 percent from 6.5 to 15.6 million in our simulation, which is roughly consistent with the change in the data — from 5.7 to 12.4 million (117 percent). The magnitude of the population inflow in Shenzhen dwarfs that of all of the other cities: the second largest inflow in absolute numbers is only 2.3 million (Shanghai), and in percentage terms, 27.8 percent (Wuhan). The special case of Shenzhen is not entirely a surprise. This city, located next to Hong Kong, was the first "Special Economic Development Zone" initiated in the 1980s, and it has always been a city in which various reforms were first tested. In the case of migration policy, the city aligns more closely with a friction-less policy as compared to other cities with similar size. In our baseline estimation, the parameter that captures the city-specific entry barriers reflects the relative ease of moving to Shenzhen. The parameter, $\delta_{\text{Shenzhen}} = 0.986$, is not only significantly lower than that of Beijing (1.06), Shanghai (1.22), or Guangzhou (1.12), it is also lower than the national average (1.00).

Most of the other cities on the top 10 list benefit either from large population inflows, such Beijing and Chongqing, or from their advantageous locations such as Wuhan, Nanjing, Tianjin and Suzhou. Tianjin, Suzhou, Hangzhou, and Foshan are only several hundred

kilometers away from cities like Beijing, Shanghai, or Shenzhen, and thus are the most direct recipient of the productivity spill overs. Wuhan and Nanjing are not next-door neighbors of the largest cities; however, they are among the most strategically placed cities in China. Wuhan sits at the crossroad of Yangtze River that connects the western inland provinces with the eastern coast, and the Beijing-Guangzhou railway, which connects the northern and the southern economic hubs of China. Similarly, Nanjing sits at the crossroad of the Yangtze river and the Great Canal, and it is also on the Beijing-Shanghai railway.

The second panel of Figure 7 plots the top 10 losers in terms of welfare. The top"loser" is Xinzhou, followed by Lvliang and Yulin, which are all located in the remote western frontier of China proper. All of the cities on the list share two common features: they lose population to large coastal cities in the east, and their remote locations denied them easy access to the productivity spillovers from the east. The same two mechanisms apply to all of the losing cities with different magnitudes. However, note that when interpreting the negative impacts on the origin cities we have omitted remittance from migrants to home due to the lack of data at the city-level. This implies that the negative impacts in our paper are closer to their upper bounds, and in the real world those impacts are partially compensated by the inflow of remittance.

4.2 Destination-Specific Migration Barriers and the Optimal City Size

Our baseline model estimates the destination-specific migration barriers for the four largest cities in China: Beijing, Shanghai, Guangzhou, and Shenzhen. These barriers to enter certain cities are used to capture the local policies designed to deter or encourage immigration. In this section, we quantify the impacts of these barriers, and discuss the optimal policy and the associated city size of these cities.

Our baseline results confirm that three out of the four cities, Beijing, Shanghai, and Guangzhou (hereafter "BSG") indeed have entry barriers higher than the national average of $\bar{\lambda}$, while the barrier to enter Shenzhen is slightly below the national average at 99 percent of its value. The barrier to move to Shanghai is the highest at about 122 percent of the

national average, followed by Guangzhou (120 percent) and Beijing (106 percent). The impacts of only 6 percent "excessive entry barrier" into Beijing turn out to be large: as we show later, removing the "excessive entry barrier" by setting it equal to national average leads to 23 percent population increase in Beijing. In our estimation these entry barriers are essentially identified from the observed migration flows: the population inflow into the BSG is surprisingly small, given the large discrepancies in welfare between these cities and their smaller neighbors; in other words, given the relative prosperity of BSG, the population of the cities would be significantly higher than what is observed in the data, if not for the higher barriers to enter them. Similarly, in the case of Shenzhen, without any policies encouraging migration to Shenzhen, the differences in other economic variables are not sufficient to explain the surge in population between the two years in the data. The relative size of the entry barriers also speaks to the desirability of the cities as destinations: barriers into Shanghai (122 percent) being higher than into Beijing (106 percent) implies that Shanghai is a more attractive destination city among migrants than Beijing. This might be rooted in the real income differences between the two cities (in both the data and our model, Shanghai has higher per capita output than Beijing), or the fact that the regions closer to Shanghai are more populous than those surrounding Beijing, leading to a larger base of potential migrants.

	Beijing	Shanghai	Guangzhou	Shenzhen
Baseline Barrier ($\delta_{(\cdot)}$)	1.06	1.22	1.12	0.99
Baseline Population (Mil.)	12.61	16.80	9.53	15.57
Counterfactural Population (Mil.)	15.57	22.62	19.83	8.83
Change in Population	23.48%	34.60%	108.09%	-43.27%
Change in Real Wage	14.38%	18.82%	48.12%	-25.90%
Change in Congestion	108.02%	180.70%	1174.67%	-86.04%
Change in Local Welfare	7.39%	-6.87%	18.86%	-18.31%
Change in National Welfare	0.96%	0.72%	3.57%	-2.54%

Table 5: Removing City-Specific Entry Barriers

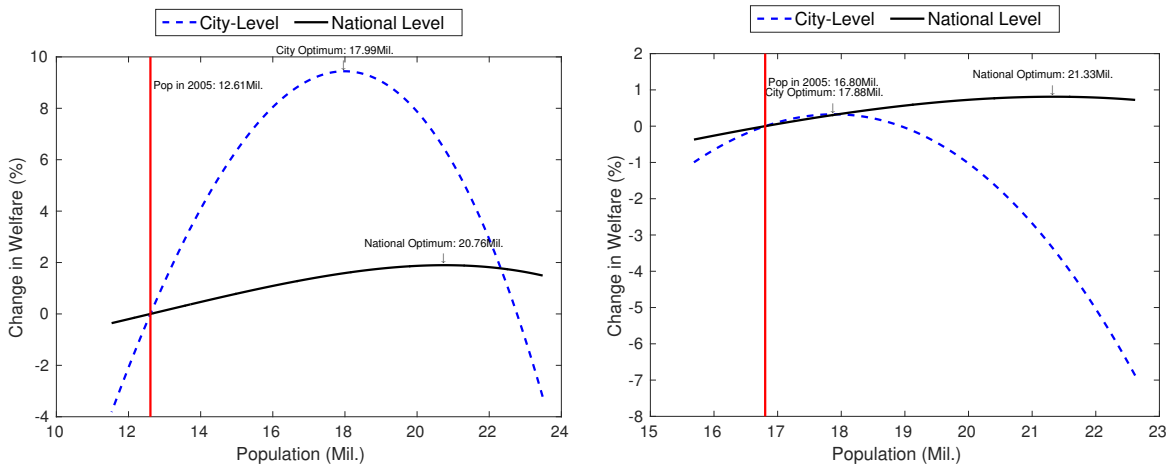
Note: This table reports the results of a counter-factual simulation in which we remove the destination-specific entry barriers by setting $\delta_{(\cdot)}$ to the national average level of 1. All of the changes are reported as percentage changes.

We evaluate the impacts of these local barriers with a series of counter-factual simulations in which we set the $\delta_{(\cdot)}$ to 1, the national average level, while keeping all of the other

parameters the same as in the baseline. The results of these simulations are reported in Table 5. In the case of Beijing, Shanghai, and Guangzhou, removing the entry barriers would lead to population increases of between 23.5 and 108.1 percent population increase. The large population inflows would lead to both higher real wage and congestion disutility. In the case of Beijing, half of the change in real income will be offset by higher congestion, resulting in a local welfare increase of around 7.4 percent; in the case of Guangzhou, around 61 percent of economic growth would be offset by congestion, resulting in a local welfare increase of around 18.9 percent. The population of Shanghai would surge to over 22 million if all of the barriers to entry were removed. The resulting burden of congestion disutility leads to a *loss of* local welfare in Shanghai of nearly 7 percent. These quantitative results show that popular support for higher entry barriers among the native populations of the largest cities can be justified based on concerns of overpopulation. In the last row of the table, we report the change in national welfare. In all three cases, the national welfare increases by around 0.7 to 3.6 percent. The case of Shanghai reveals the potential conflict between the local and national welfare: although the natives living in the city suffer a 7 percent welfare loss, the entire nation gains around 0.7 percent in welfare, mainly because of the productivity spillovers from Shanghai. We discuss this conflict in more detail when we compute the optimal city sizes.

Shenzhen's local entry barrier is smaller than the national average, and thus, in our counter-factual simulation, Shenzhen increases the relative difficulty of entry as compared to the baseline. This leads to around a 43-percent loss in population, and around an 18-percent loss in local welfare. The large response of the population flow to a relatively small change in local entry barriers is mainly due to the special location of Shenzhen: it sits within a greater metropolitan area, the Pearl River Delta, which contains many large cities with comparable levels of development, such as Guangzhou, Zhuhai and Dongguan. Workers residing within this cluster of cities are highly mobile, as the migration frictions within the Pearl River Delta are low and the relative differences in real wages across cities are small. As a result, a small change in entry barriers can easily affect the migration decisions of many workers.

The destination-specific migration barrier, $\delta_{(.)}$, is often determined by the local government. We use our model to quantify the "optimal" level of entry barrier at both the national



(a) Beijing (b) Shanghai

Figure 8: Optimal City Size of Beijing and Shanghai

Note: This figure plots the local and national welfare of Beijing and Shanghai as a function of population. The change in population is driven by the local entry barrier, δ_i. The welfare in the benchmark model (the red solid line) is normalized to 1.

and city levels. To do this for a given city i, we simulate the model with different values of δ_i within a certain interval while keeping all of the other parameters fixed. We then proceed to determine the value of δ_i and the corresponding population of the city that maximizes its own welfare and the national welfare. The results for the case of Beijing are shown in Figure 8. In the figure, we plot the percentage change in welfare against the population, which is a function of δ_i, in Beijing. We find that the existence of entry barriers significantly lowers welfare at both the local and national levels. The population that optimizes the welfare in Beijing should be around 18 million, which is around 50 percent higher than the population in Beijing as of 2005 (12.61 million). If Beijing were to adopt the optimal entry barrier, the gain in the real wage would outweigh the loss in congestion disutility, leading to a welfare gain of around 9.44 percent. Interestingly, the population that maximizes the local welfare in Beijing is much smaller than the one that maximizes the national welfare, which is around 22 million. This implies that the positive spillovers through intercity trade still dominate the congestion disutility within the city. From the perspective of the central government, the local citizens of Beijing would bear a heavier burden from congestion for the greater welfare of the entire country. Differences between the optimal local and national policies can be consistently found for all of the other cities that we have analyzed (see Figure

12 in the appendix), which highlights the conflict of interest between the different levels of government. As of 2015, the population of Beijing had already exceeded 22 million, which is slightly over-populated even by the national optimum, and over-populated by as much as $22/18 \approx 22$ percent by the local optimum in 2005.

Despite the fact that Shanghai has the highest entry barrier (1.22), we find that at around 17.8 million, the size of the city (16.8 million) was already close to the local optimum as of 2005. Further lowering the entry barriers would actually lead to sub-optimal results at the local level, as the surge in congestion would outweigh the gain in real wage. Nevertheless, from a national perspective, Shanghai would still expand by around $21.33/16.8 - 1 \approx 27$ percent.

4.3 Internal and International Trade Liberalization

In this section, we study the impacts of internal and international trade liberalization and their interactions with internal migration in China.

Internal Trade Liberalization. We simulate a counter-factual world in which the internal trade frictions, $\bar{\tau}$, are lowered by 10 percent, and compare the results to our baseline model. To single out the effects of internal trade liberalization, we increase the international trade barrier, τ_{row}, by 11.1 percent, so that the effective trade barrier between China and the ROW, $\bar{\tau} \cdot \tau_{\text{row}}$, is the same between the counter-factual and the benchmark model. All of the other parameters are the same as in the benchmark simulation.

The results are reported in the four panels in Figure 9. Lowering the internal trade barrier by 10 percent increases the aggregate real income by around 4.5 percent, as shown in Panel (a). The entire country can benefit from internal trade liberalization, mainly through two channels: 1) the direct benefit in terms of reduced transportation costs, which is the same as the "gains from trade" as in a standard trade model, and 2) the indirect benefit of trade-induced migration as detailed in the previous section. To quantify the relative importance of the two channels, we run another counter-factual simulation in which we reduce the internal trade barriers while shutting down the migration channel.[16] Internal migration plays a

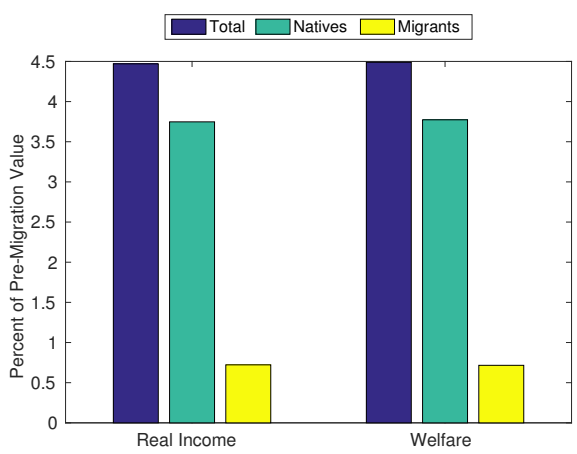
[16] We set $\bar{\lambda}$ to a sufficiently high number, and start the model using the equilibrium population distribution

relatively minor role in amplifying the gains from internal trade liberalization: without any migration, the overall gain in the real wage is 4.1 percent, as shown in Panel (b) of the same figure. This implies that $4.1/4.5 \approx 91.1$ percent of the gain is through the first channel, while the other 8.9 percent is the amplification from trade-induced migration. The relative insignificance of the amplification effects is probably due to the small scale of the trade-induced migration flow: when we reduce $\bar{\tau}$ by 10 percent, the aggregate stay rate drops slightly from 82.7 to 82.2 percent.

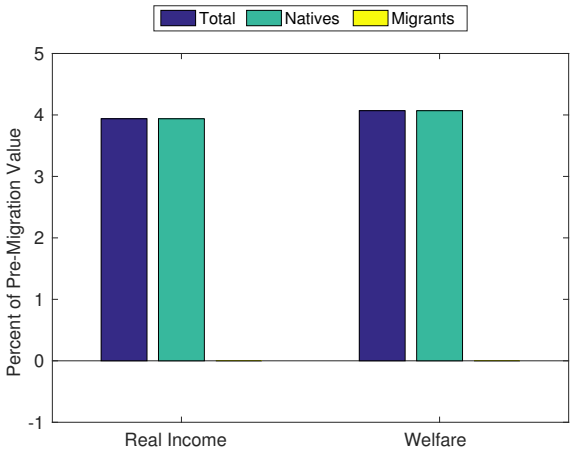
The insensitivity of internal migration to internal trade frictions is due to the spatial distribution of the "gains from trade," as shown in the third and the fourth panels of Figure 9. While all of the cities benefit from internal trade liberalizations, the small and inland cities gain much more than the large and coastal ones. For example, as shown in Panel (d) of the figure, the change in real income in cities such as Beijing and Shenzhen is only around 2.6 percent, while the smaller cities can enjoy gains at around 9.0 percent. This is an expected result from trade models following the works of Krugman [1980]: small economies usually benefit more from trade liberalization because after liberalization, the number of new imported varieties relative to the existing market size is larger in smaller cities, leading to a steeper drop in the ideal price index. In other words, internal trade liberalization tends to narrow down the gaps in real wages across space: the spatial inequality measured by the coefficient of variation of the real wage across cities is 0.48 in the baseline model and 0.47 in the counter-factual. As a result, the need to migrate to large cities is mitigated in the first place.

The fact that small cities benefit more than large ones also implies that internal trade liberalization would not lead to a sharp increase in congestion disutility. Panel (a) in Figure 9 shows that this is indeed the case: the aggregate welfare gain is around 4.47 percent, which is effectively the same as the aggregate income gain, indicating the negligible change in congestion-disutility. Compared to the case of reductions in migration frictions in the previous section, where around 17.8 percent of the gain in real income was offset by higher congestion disutility, our results suggest that reductions in internal trade frictions seem to

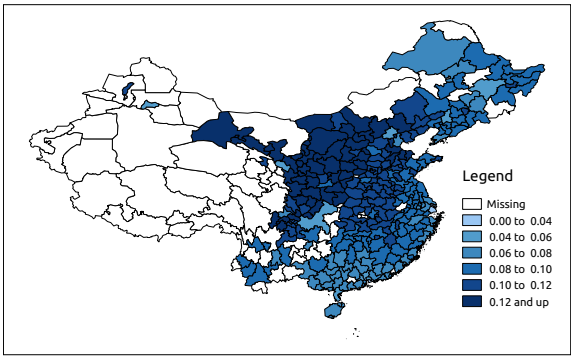
from the baseline model as the initial distribution. We then set $\bar{\tau}$ and τ_{row} to be the same in the internal-trade-reduction counter-factual to simulate the results reported in Panel (b) of Figure 9.



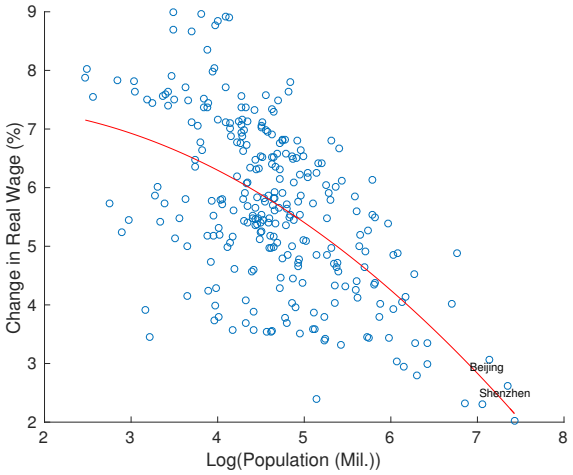
(a) Aggregate Impacts with Migration



(b) Aggregate Impacts without Migration



(c) Spatial Distributions of Welfare Gains



(d) Initial Population and Change in the Real Wage

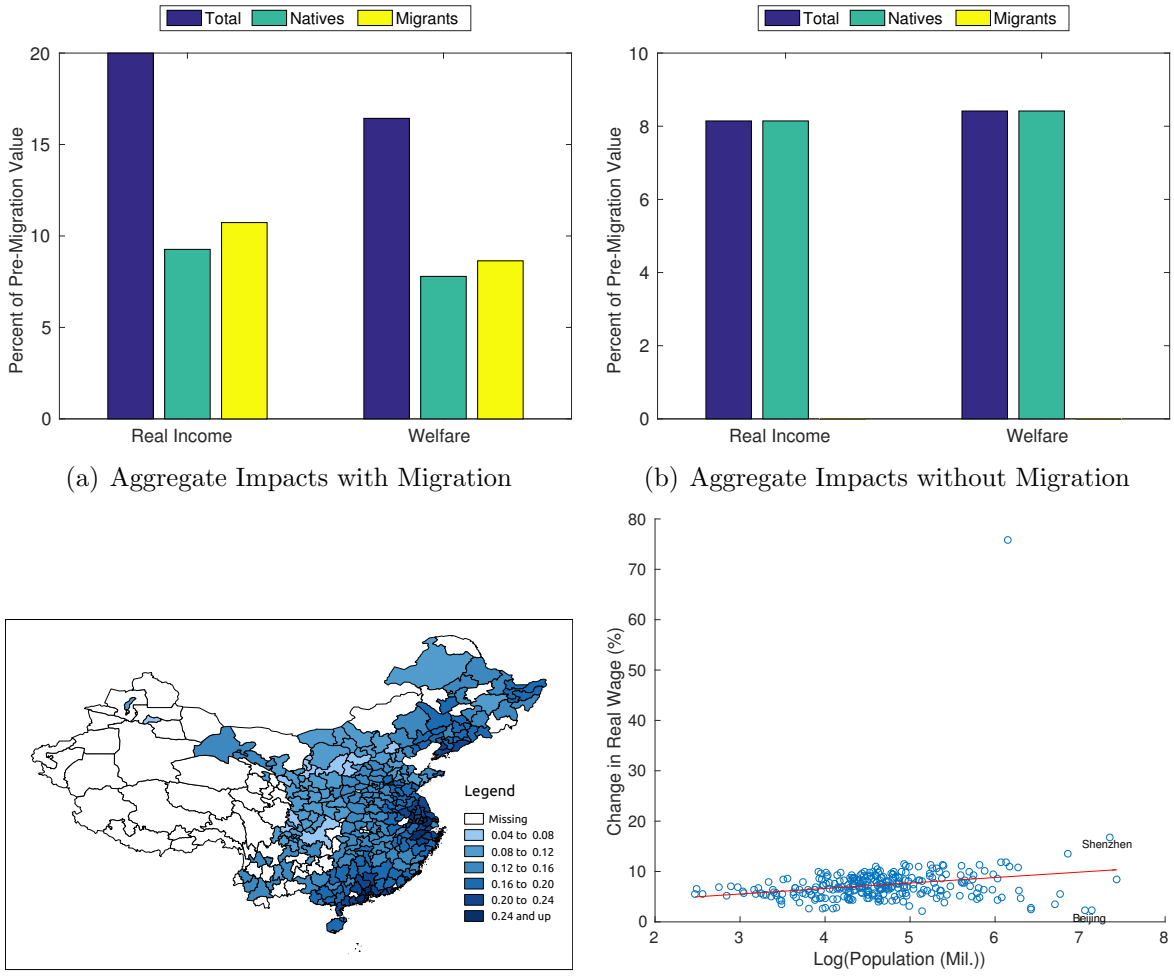
Figure 9: Aggregate Impacts of Internal Trade Liberalization

Note: The figures report the aggregate impacts and the direction of population flows from lowering the internal trade barrier, $\bar{\tau}$, by 10 percent while keeping $\bar{\tau} \cdot \tau_{\text{row}}$ the same as in the benchmark model.

be more "efficient".

International Trade Liberalization Similar to our previous exercise, we study the impacts of the international trade barrier by simulating a counter-factual world in which τ_{row} is lowered by 10 percent, while all of the other parameters are kept the same as in the baseline model. The results are reported in Figure 10 in a similar manner as in the previous section.

A 10-percent reduction in the international trade barrier leads to a 20.0-percent increase in real income. The trade-induced migration plays a much more important role than in the case of internal trade liberalization. The reduction in trade barriers in this case directly



(a) Aggregate Impacts with Migration

(b) Aggregate Impacts without Migration

(c) Spatial Distribution of Welfare Gain

(d) Initial Population and Change in Real Wage

Figure 10: Aggregate Impacts of International Trade Liberalization

Note: The figures report the aggregate impacts and the direction of population flows of lowering the international trade barrier, τ_{row}, by 10 percent while keeping all of the other parameters the same as in the benchmark model.

lowers the aggregate stay rate from 83 percent in the benchmark to 76 percent. Without migration, a 10-percent reduction in trade barriers will only lead to 8.1 percent growth in real income, as shown in the second panel of the graph. This implies that intercity migration can amplify the gains from trade as computed in a standard trade model without migration by as much as $20.0/8.1 - 1 \approx 147$ percent. The amplification mechanism mainly works through the impacts of migration on local wage rates. Reductions in trade barriers usually lead to the rapid expansion of the firms located along the coasts. However, without intercity migration, higher labor demand quickly pushes up the wage rates, which in turn increases

the marginal costs of production in those cities. This effectively limits the extent to which firms can grow in the coastal cities, and thus dampen the gains from trade. In contrast, once intercity migration is allowed, higher wage rates in the coastal cities draw the workers from inland cities, pushing out the labor supply curve, and dampen the equilibrium wage rate in these cities. This enables the exporting firms in the coastal cities to grow much larger relative to the scenarios without migration, and eventually leads to higher gains from trade.

Through the lens of international trade models, almost all of the gains from trade come from the reallocation of resources within the country: the neo-classical trade models usually emphasize inter-industry reallocation dictated by comparative advantage, and the new trade models following the works of Melitz [2003] highlight the gains due to cross-firm reallocation. Our work here shows that the spatial reallocation of production factors, a previously overlooked channel, quantitatively dominates the traditional channels. If we allow workers to migrate to regions with better access to the international market, the impacts of equilibrium wage rate, which usually works to limit the gains from trade, can be greatly mitigated. The resulting amplification effects can generate a gain from trade to be more than twice of those measured in standard trade models without migration. The spatial reallocation component also pushes our model outside of the definition of Standard Trade Models as in Arkolakis et al. [2012], which implies that the amplification effects of migration cannot be simply captured by the overall openness of the country and the trade elasticity as shown in Arkolakis et al. [2012].

Trade-induced migration is generally directed toward large coastal cities, as those cities usually benefit the most from reductions in international trade barriers. Panels (c) and (d) in Figure 10 highlight this pattern of welfare and real income gain. Despite the amplification effects of trade-induced migration, higher concentration of population among the coastal cities toll on the national welfare in terms of higher congestion disutility. National welfare only increase by 16.4, implying that around $1 - 16.4/20.0 \approx 18.0$ percent of the gain in the real wage is offset by higher congestion disutility. This is a sharp contrast to the case of internal trade liberalization where the congestion disutility is virtually unchanged.

The gains from international trade liberalization in our model are higher than what is

usually seen in the quantitative literature.[17] We decompose the gains from international trade into different channels in Table 6 to highlight the relative importance of the new elements in our model. Shutting down the migration channel reduces the gains from trade from 20 percent to 8.14 percent, and shutting down the firm entry reduces the gain to 5.91 percent. If we shut down both channels the gains from trade in our framework is comparable to the majority of estimates in the literature, at around 5.39 percent. The fact that internal migration channel has a large impact on gains from trade is mainly due to its interaction with firm entry. Without firm entry, the gains from migration is significantly reduced (5.91 percent) compared to the benchmark exercise. Workers no longer anticipate that more varieties and lower prices will be offered in the destination cities, and thus fewer workers choose to migrate and the gains from trade become limited. For example, with firm entry the reductions in τ_{row} induces a 21 percent population growth in Shenzhen; However without firm entry, the same reduction in τ_{row} only increases the population in Shenzhen by 2.9 percent. The strong interaction between firm entry and migration also explains why shutting down both channels has similar effect as compared to only shutting down the entry channel (5.39 percent v.s. 5.91 percent): without firm entry the migration flow is already tiny, and thus further banning migration will not lead to a sizable change in real income. The second column reports the gains from trade using welfare, instead of real wage, and arrives at similar results. Note that under "no migration" the percentage gains in real wage and welfare are slightly different, because congestion costs enter the utility function additively, not multiplicatively in our model.

5 Conclusion

In this paper, we develop a multi-city general equilibrium framework with endogenous firm dynamics and a migration decision. We then structurally estimate the model to quantify the welfare implications of migration and trade frictions at both the aggregate and local levels. The model is able to rationalize about 22 percent of the GDP growth between the 2000 and

[17]For example, the closest counter-part to our model, di Giovanni and Levchenko [2013], under a similar quantification with fat-tailed firm size distribution, found that a 10 percent reduction in variable trade barriers on average increases welfare by 4.3 percent.

10% Reduction in τ_{row}	$\Delta \log$(Real Wage)	$\Delta \log$(Welfare)
Benchmark	0.2000	0.1643
No Migration	0.0814	0.0842
No Firm Entry	0.0591	0.0587
No Migration and Firm Entry	0.0539	0.0557

Table 6: Decomposing the Gains from Trade

Note: The table decomposes the gains from lowering international trade frictions by 10% into two channels: migration and firm entry. In "No Migration" and "No Firm Entry" setting, the population and firm distributions are set to be the same as "Benchmark" prior to the reduction in τ_{row}.

2005 by simply reallocating labor across the country. Moreover, conflicting interests prevail between the local and central government in setting the destination-specific entry barrier. In general, the central government prefers a lower entry barrier than the local government to induce greater agglomeration effects but at the expense of imposing heavy congestion disutility on local residents. Finally, internal trade liberalization tends to reduce spatial income inequality, whereas international trade liberalization tends to induce more migration to richer coastal area.

An explicitly modeled agricultural sector is absent in the current setup. Moreover, the cities in our model are all ex-ante identical except for initial population size. A further extension could be made to incorporate an agricultural sector with city-specific land endowment as an input of production. This would enable us to distinguish between rural-urban and urban-urban migration. Furthermore, it would also be worthwhile to explore a dynamic model with an endogenous migration decision.

References

Allen, Treb and Costas Arkolakis, "Trade and the Topography of the Spatial Economy," *The Quarterly Journal of Economics*, 2014, *1085*, 1139.

Anderson, James E. and Eric van Wincoop, "Gravity with Gravitas: A Solution to the Border Puzzle," *American Economic Review*, March 2003, *93* (1), 170–192.

Arkolakis, Costas, Arnaud Costinot, and Andres Rodriguez-Clare, "New Trade Models, Same Old Gains?," *American Economic Review*, February 2012, *102* (1), 94–130.

Au, Chun-Chung and J. Vernon Henderson, "Are Chinese Cities Too Small?," *Review of Economic Studies*, 2006, *73* (3), 549–576.

Axtell, Robert L, "Zipf Distribution of U.S. Firm Sizes," *Science*, 2001, *293* (5536), 1818–1820.

Brandt, Loren, Chang-Tai Hsieh, and Xiaodong Zhu, "Growth and structural transformation in China," *Chinas great economic transformation*, 2008, pp. 683–728.

Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro, "Trade and labor market dynamics," *NBER Working Paper*, 2015, *21149.*

Chan, Kam Wing; Peter Bellwood, "China, Internal Migration," in "The Encyclopedia of Global Migration," Blackwell Publishing, 2011.

Chow, Gregory C, "Capital formation and economic growth in China," *The Quarterly Journal of Economics*, 1993, pp. 809–842.

di Giovanni, Julian and Andrei A. Levchenko, "Country size, international trade, and aggregate fluctuations in granular economies," *Journal of Political Economy*, 2012, *120* (6), 1083–1132.

_ and _ , "Firm entry, trade, and welfare in Zipf's world," *Journal of International Economics*, 2013, *89* (2), 283–296.

_ , _ , and Francesc Ortega, "A Global View Of Cross-Border Migration," *Journal of the European Economic Association*, 02 2015, *13* (1), 168–202.

Eaton, Jonathan and Samuel Kortum, "Technology, Geography, and Trade," *Econometrica*, Sep. 2002, *70* (5), 1741–1779.

_ , _ , and Francis Kramarz, "An Anatomy of International Trade: Evidence From French Firms," *Econometrica*, 09 2011, *79* (5), 1453–1498.

Fan, Jingting, "Internal Geography, Labor Mobility, and the Distributional Impacts of Trade," *Labor Mobility, and the Distributional Impacts of Trade (January 2015)*, 2015.

Grogger, Jeffrey and Gordon H. Hanson, "Income maximization and the selection and sorting of international migrants," *Journal of Development Economics*, 2011, *95* (1), 42 – 57. Symposium on Globalization and Brain Drain.

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

Krugman, Paul, "Scale Economies, Product Differentiation, and the Pattern of Trade," *American Economic Review*, December 1980, *70* (5), 950–59.

McCallum, John, "National Borders Matter: Canada-U.S. Regional Trade Patterns," *American Economic Review*, June 1995, *85* (3), 615–23.

McFadden, Daniel, "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, September 1989, *57* (5), 995–1026.

_ **and Paul A Ruud**, "Estimation by Simulation," *The Review of Economics and Statistics*, November 1994, *76* (4), 591–608.

Melitz, M.J., "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 2003, *71* (6), 1695–1725.

Ortega, Francesc and Giovanni Peri, "The Effect of Income and Immigration Policies on International Migration," *Migration Studies*, 2013, *1* (1), 47–74.

_ **and** _ , "Openness and Income: The Roles of Trade and Migration," *Journal of International Economics*, 2014, *92* (2), 231–251.

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti, "Growing Like China," *American Economic Review*, February 2011, *101* (1), 196–233.

Tombe, Trevor and Xiaodong Zhu, "Trade, Migration and Productivity: A Quantitative Analysis of China," 2015. working paper.

World Bank, "China - Investment Climate Survey 2005," Technical Report, Enterprise Analysis Unit, World Bank Group 2005.

A Tables and Figures

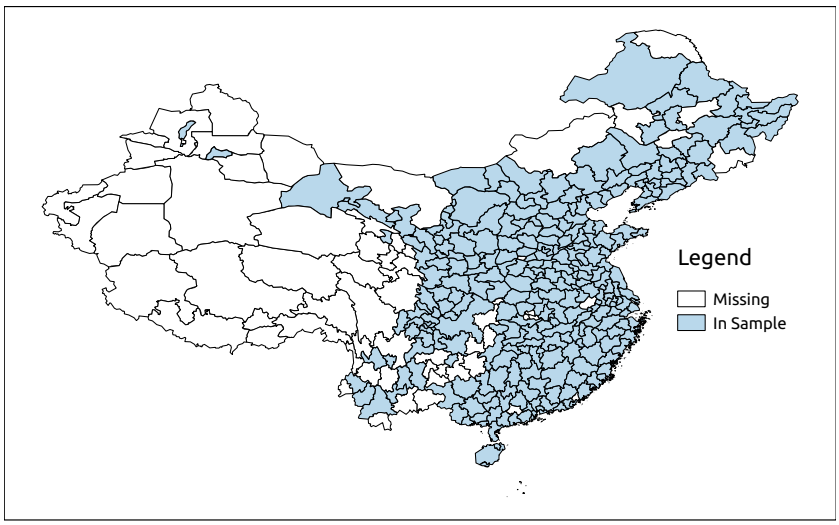
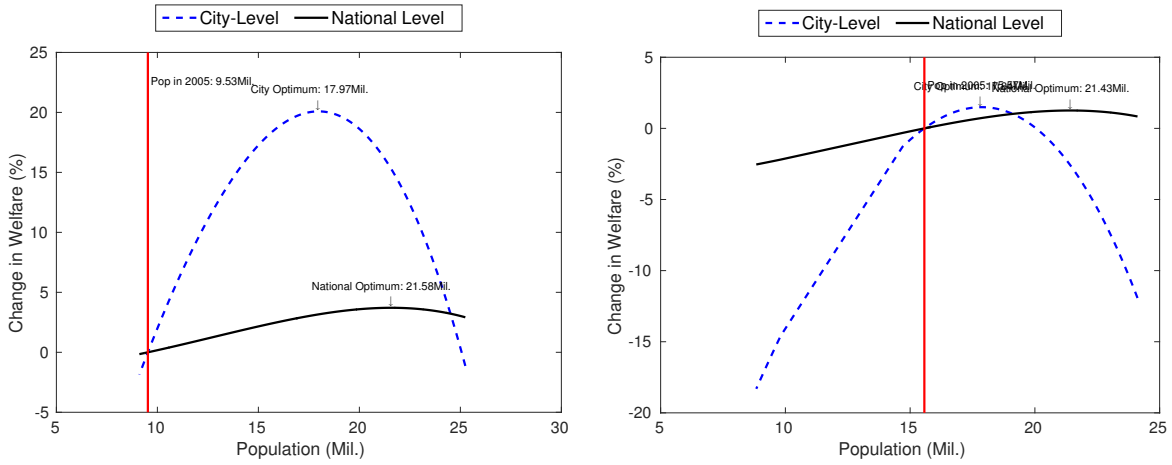


Figure 11: Prefecture-level Chinese Cities

Note: This graph shows the 279 prefecture-level cities included in our sample. All of the cities that are included appear both in the Chinese Statistical Yearbooks and the 2005 micro survey.



(a) Guangzhou (b) Shenzhen

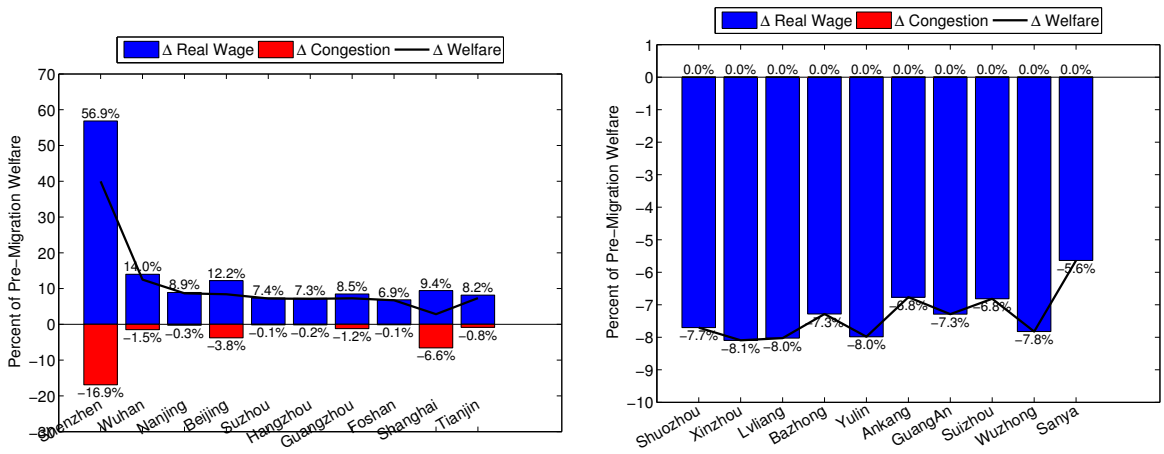
Figure 12: Optimal City Size of Guangzhou and Shenzhen

Note: This figure plots the local and national welfare of Guangzhou and Shenzhen as a function of population. The change in population is driven by the local entry barrier, δ_i. The welfare in the benchmark model (the red solid line) is normalized to 1.

Name	Data	Model	Diff.
Num. Firms	8.441	8.834	4.7%
Trade Share	0.625	0.681	9.0%
Tail Index	1.030	1.035	0.4%
Stay Rate	0.881	0.827	-6.1%
Std. Stay Rate	0.090	0.076	-15.3%
Corr(log(pop) inflow)	0.360	0.385	7.1%
Stay Rate Top 10	0.980	0.982	0.3%
Stay Rate Top 20	0.970	0.970	0.0%
Stay Rate Top 40	0.947	0.942	-0.5%
Stay Rate Other	0.869	0.807	-7.2%
Std(SR) Top 10	0.019	0.016	-16.5%
Std(SR) Top 20	0.020	0.023	19.2%
Std(SR) Top 40	0.039	0.038	-1.3%
Std(SR) Other	0.091	0.062	-32.4%
(Export+Import)/GDP	0.594	0.565	-4.8%
ROW/China Size	21.320	21.486	0.8%
Inflow Rate Beijing	0.166	0.165	-0.3%
Inflow Rate Shanghai	0.138	0.138	-0.3%
Inflow Rate Guangzhou	0.151	0.151	-0.3%
Inflow Rate Shenzhen	0.539	0.584	8.3%

Table 7: Model Fit: Targeted Moments

Note: The table reports all of the targeted moments in our SMM, the moments in the data, and their counter parts in the model. For more details on the moments and the data source, see the main text. The third column reports the differences between the data and the model in percentage points.



(a) Top 10 Inflow

(b) Top 10 Outflow

Figure 13: Welfare from Migration, by Migration Flow

Note: The two panels above plot the change in the real wage and congestion costs as a percentage of welfare before and after migration. The first panel plots the top 10 cities in terms of the population inflow rate, and the second panel plots the bottom 10 cities by the same measure.

B Solving the Model

In each city, we need to solve $\left\{w_j, I_j^N, I_j^T, P_j^T, P_j^N\right\}$. When we assume that $\frac{1}{a}$ follows a type-1 Pareto distribution, $\Pr\left(\frac{1}{a} < y\right) = 1 - \left(\frac{b}{y}\right)^{\theta}$, then we can explicitly derive

$$\begin{aligned} P_i^N &= \frac{1}{b_N}\left[\frac{\theta_N}{\theta_N - (\varepsilon_N - 1)}\right]^{\frac{-1}{\theta_N}} \frac{\varepsilon_N}{\varepsilon_N - 1}\left(\frac{X_i^N}{\varepsilon_N}\right)^{-\frac{\theta_N - (\varepsilon_N - 1)}{\theta_N(\varepsilon_N - 1)}} \\ &\quad \left(I_i^N\left(\frac{1}{c_i^N}\right)^{\theta_N}\left(\frac{1}{c_i^N f_{ii}^N}\right)^{\frac{\theta_N - (\varepsilon_N - 1)}{\varepsilon_N - 1}}\right)^{\frac{-1}{\theta_N}} \qquad (8) \\ P_i^T &= \frac{1}{b_T}\left[\frac{\theta_T}{\theta_T - (\varepsilon_T - 1)}\right]^{\frac{-1}{\theta_T}} \frac{\varepsilon_T}{\varepsilon_T - 1}\left(\frac{X_i^T}{\varepsilon_T}\right)^{-\frac{\theta_T - (\varepsilon_T - 1)}{\theta_T(\varepsilon_T - 1)}} \\ &\quad \left(\sum_{j=1}^{J} I_j^T\left(\frac{1}{\tau_{ij} c_j^T}\right)^{\theta_T}\left(\frac{1}{c_j^T f_{ij}^T}\right)^{\frac{\theta_T - (\varepsilon_T - 1)}{\varepsilon_T - 1}}\right)^{\frac{-1}{\theta_T}} \qquad (9) \end{aligned}$$

The free entry condition in each city/sector gives:

$$\begin{aligned} \frac{X_j^N}{\varepsilon P_{jN}^{1-\varepsilon}}\left(\frac{\varepsilon}{\varepsilon - 1} c_j^N\right)^{1-\varepsilon} \frac{\theta b^{\theta} a_{jN}^{1+\theta-\varepsilon}}{1 + \theta - \varepsilon} &= c_j^N f_e + b^{\theta} a_{jN}^{\theta} c_j^N f_{jj} \qquad (10) \\ \sum_{i=1}^{J} \frac{X_i^T}{\varepsilon P_{iT}^{1-\varepsilon}}\left(\frac{\varepsilon}{\varepsilon - 1}\tau_{ij} c_j^T\right)^{1-\varepsilon} \frac{\theta b^{\theta} a_{ij}^{1+\theta-\varepsilon}}{1 + \theta - \varepsilon} &= c_j^T f_e + \sum_{i=1}^{C} b^{\theta} a_{ij}^{\theta} c_j^T f_{ij} \qquad (11) \end{aligned}$$

where a_{jN} and a_{ij} are from zero-profit conditions:

$$\begin{aligned} a_{jN} &= \frac{\varepsilon - 1}{\varepsilon}\frac{P_{jN}}{c_j^N}\left(\frac{X_j^N}{\varepsilon c_j^N f_{jj}}\right)^{\frac{1}{\varepsilon - 1}} \\ a_{ij} &= \frac{\varepsilon - 1}{\varepsilon}\frac{P_{iT}}{\tau_{ij} c_j^T}\left(\frac{X_i^T}{\varepsilon c_j^T f_{ij}}\right)^{\frac{1}{\varepsilon - 1}} \end{aligned}$$

where

$$\begin{aligned} X_i^N &= \alpha w_i L_i + (1 - \beta_N)\,\eta_N X_i^N + (1 - \beta_T)\,\eta_T X_i^T \\ X_i^T &= (1 - \alpha)\,w_i L_i + (1 - \beta_N)\,(1 - \eta_N)\,X_i^N + (1 - \beta_T)\,(1 - \eta_T)\,X_i^T \end{aligned}$$

Finally, the balance of trade condition requires $X_i^T = \sum_{j=1}^J X_{ji}^T$, so in each country i we have $w_i L_i$ equal to the following:

$$w_i L_i = \sum_{j=1}^{J} \frac{I_i^T \tau_{ji}^{-\theta_T} \left(f_{ji}^T\right)^{-\frac{\theta_T-(\varepsilon_T-1)}{\varepsilon_T-1}} \left(w_i^{\beta_T}\left[\left(P_i^N\right)^{\eta_s}\left(P_i^T\right)^{1-\eta_T}\right]^{1-\beta_T}\right)^{-\frac{\theta_T-(\varepsilon_T-1)}{\varepsilon_T-1}}}{\sum_{l=1}^J I_l^T \tau_{jl}^{-\theta_T} \left(f_{jl}^T\right)^{-\frac{\theta_T-(\varepsilon_T-1)}{\varepsilon_T-1}} \left(w_l^{\beta_T}\left[\left(P_l^N\right)^{\eta_s}\left(P_l^T\right)^{1-\eta_T}\right]^{1-\beta_T}\right)^{-\frac{\theta_T-(\varepsilon_T-1)}{\varepsilon_T-1}}} w_j L_j \quad (12)$$

Given the initial distribution of labor $\left\{\overline{L_i}\right\}$, the labor supply in country i is $L_i^S = \sum_{j=1}^{J} m_{ij}\overline{L_j}$

Then, the total demand for labor in city i is

$$\begin{aligned} L_i^D &= I_i^N \frac{\beta^N C_i^N}{w_i} \frac{X_i^N}{\left(P_i^N\right)^{1-\varepsilon}} \left[\frac{\varepsilon}{\varepsilon-1}\tau_{ii} c_i^N\right]^{-\varepsilon} \frac{\theta b^\theta \left(a_{ii}^N\right)^{1+\theta-\varepsilon}}{1+\theta-\varepsilon} \\ &\quad + I_i^T \sum_{j=1}^{J} \frac{\beta^T C_i^T}{w_i} \frac{X_j^T}{\left(P_j^T\right)^{1-\varepsilon}} \left[\frac{\varepsilon}{\varepsilon-1}\tau_{ji} c_i^T\right]^{-\varepsilon} \frac{\theta b^\theta \left(a_{ji}^T\right)^{1+\theta-\varepsilon}}{1+\theta-\varepsilon} \end{aligned}$$

Labor market clearing gives

$$\sum_{j=1}^{J} m_{ij}\overline{L_j} = I_i^N \frac{\beta^N C_i^N}{w_i} \frac{X_i^N}{\left(P_i^N\right)^{1-\varepsilon}} \left[\frac{\varepsilon}{\varepsilon-1}\tau_{ii} c_i^N\right]^{-\varepsilon} \frac{\theta b^\theta \left(a_{ii}^N\right)^{1+\theta-\varepsilon}}{1+\theta-\varepsilon}$$

$$+ I_i^T \sum_{j=1}^{J} \frac{\beta^T C_i^T}{w_i} \frac{X_j^T}{\left(P_j^T\right)^{1-\varepsilon}} \left[\frac{\varepsilon}{\varepsilon-1}\tau_{ji} c_i^T\right]^{-\varepsilon} \frac{\theta b^\theta \left(a_{ji}^T\right)^{1+\theta-\varepsilon}}{1+\theta-\varepsilon}, \forall i$$

In total, we have $5 * J + (J-1)$ equations to solve for $w_j, I_j^N, I_j^T, P_j^T, P_j^N, L_j$, where we normalize one city's wage rate to be 1.

C Numerical Implementation

C.1 Structural Estimation

In this appendix, we provide the details on how to estimate the structural parameters of the model, or equivalently, to solve the minimization problem stated in Equation (6). Our entire algorithm consists of two layers:

A. **The Inner Layer.** The inner layer of the algorithm solves the model conditional on all inputs, including the parameter of interest, Θ. The equilibrium conditions of the model are a large system of non-linear equations. We solve the system with a standard nested-loops algorithm:

Step 1. Start with an initial guess of the equilibrium population distribution. Conditional on the guess, solve for the equilibrium number of entrants and operating firms in each sector and city, product prices, and wage rates in each city.

Step 2. Conditional on the equilibrium results solved in the previous step, compute the bilateral migration matrix and the implied equilibrium population distribution.

Step 3. Compare the initial guess with the implied population distribution. If the differences are below a certain threshold, exit the algorithm; otherwise, update the initial guess with the implied distribution and iterate back to step 1.

B. **The Outer Layer.** The outer layer of the algorithm solves the minimization problem conditional on the solutions provided in the inner layer. Conditional on an input vector Θ, the inner layer finds the distance between the model and the data moments; the outer layer will try to find the input vector Θ that minimizes the distance. We implement an iterative particle swarm optimization algorithm (PSO) to solve the minimization problem. At iteration t, the algorithm can be described as follows:

Step 1. Start with an initial input of the iteration, Θ_t.

Step 2. Define a subspace around Θ_t, and randomly draw n initial positions of Θ (particles) within the subspace. Denote the position of particle i as $p(i)$.

Step 3. For each particle i, define a random neighborhood particle set and denote the neighborhood set of particle i as $b(i)$.

Step 4. Evaluate the model at each of the n particles. Denote the global best solution as $g*$, and the best solution within the neighborhood of particle i as $b^*(i)$.

Step 5. Update the position of each particle i as

$$p'(i) = W_1 * p(i) + u(1) * W_2 * g^* + u(2) * W_3 * b^*(i).$$

$p'(i)$ is the new position, $p(i)$ is the old position, $u(\cdot)$ are uniformly distributed random numbers, and $W_{(\cdot)}$ are weights.

Step 6. Iterate between steps 3 and 5 until all of the particles converge to the same position, or we can no longer improve g under certain stall limits.

Step 7. Check if the best solution from the previous step is an improvement over the initial guess, Θ_t:

- If it is an improvement, reset the stall counter to 0 and update the initial guess with the current best solution, then iterate starting from step 1 again.
- If it is not an improvement, add 1 to the stall counter, and restart from step 1 with the same initial guess, but different subspace and/or random seed.

Step 8. Exit if Θ_t cannot no longer be improved (stall counter exceeds stall limit).

C.2 Bootstrapping

We estimate the standard errors using a 200-repetition bootstrapping. In each repetition, we bootstrap the following data samples to re-compute the target moment in equation 6:

1. **The 2005 Micro-Census**. This micro-census contains individual-level data. In each bootstrap, we impose strata restrictions so that the number of observations in each city equals that in the original data set. After the bootstrapping, we re-compute the bilateral migration matrix, and thus all of the moments based on the matrix.

2. **The Investment Climate Survey, 2005**. This survey is carried out at the firm-level. In the bootstrapping, we do not impose strata restrictions, and we thus directly re-draw from the entire sample. After the bootstrapping, we re-compute the the internal-trade-to-GDP ratio.

3. **The Second Economic Census, 2004**. This census is at the firm-level. We first aggregate up the count data to the city-level, and then bootstrap with the sample of 279 cities. In each bootstrap, we sort the cities and compute the number of firms in the top-20 cities in the sample. In the corresponding bootstrapping estimation, we pick the same 20 cities as in the specific bootstrapping sample to ensure consistency.

After all of the bootstrapped moments have been computed, we apply the entire two-layer algorithm for each of the 200 samples to generate the bootstrapped distribution of each estimated parameter. The standard errors of each parameter can be directly derived from the bootstrapped distribution.