1-2020

Geography, trade, and internal migration in China

Lin MA
Singapore Management University, linma@smu.edu.sg

Yang TANG

Geography, Trade, and Internal Migration in China

Lin Ma Yang Tang *

June 13, 2019

Abstract

This paper quantitatively studies the local welfare impacts of inter-city migration in China. We structurally estimate a trade model with endogenous migration decisions using data from 279 prefecture-level cities. The results suggest that inflows of migrant workers increase welfare in the destination cities between 2000 and 2005 despite their negative impacts on congestion and nominal wage. The positive local impacts of migration depend crucially on the endogenous firm entry. The positive impacts in the destination cities also spill over to the neighboring cities through inter-city trade, often leading to higher welfare gains in the nearby cities than the destination cities themselves. We also show that further relaxing the Hukou restrictions in the largest Chinese cities is welfare-improving to the local residents.

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 the editor, Nathaniel Baum-Snow, and two anonymous referees for their suggestion and comments. We also 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. Zhongchen Hu provided excellent research assistance. Yang Tang acknowledges financial support from Singapore MOE-Tier 1 research fund(M4011853). We are solely responsible for the remaining errors.

1 Introduction

Over the past several decades, China has witnessed the largest wave of migration in history. Following the relaxation of the household registration (Hukou) system in the post-Mao era, around 340 million individuals (Chan, 2011) — roughly the entire population of the U.S. — have traveled thousands of miles away from their hometowns in search for a better life. Migration at this scale is bound to stir up controversy, and at the center of the debate lies the question of the local impacts: do the residents in the destination cities benefit from the influx of migrants? The local impacts matter because if the people in the receiving cities suffer from immigration, then further reforms on labor mobility will be politically divisive and morally contestable regardless of the potential gains at the national level. Unfortunately, the local impacts of immigration are often unclear. On the one hand, migrants bring prosperity to their new homes as they introduce fresh entrepreneurial ideas and increase labor supply. On the other hand, the new comers also compete with the locals for jobs, contribute to congestion, and push up the housing prices. In this paper, we study the local impacts of migration through the lens of a general equilibrium framework. We show that quantitatively, the benefits of migrants outweigh the detriments in the destination cities, and thus further liberation of migration barriers in China will improve not only the aggregate productivity but also the local welfare in the receiving cities.

To analyze the local impacts of migration, we introduce endogenous migration decisions into a quantitative trade model along the lines of Melitz (2003), Eaton et al. (2011), and di Giovanni and Levchenko (2012). In our model, individuals choose locations depending on the real wage and congestion dis-utility, origin-destination specific migration barriers, and idiosyncratic preferences. The real wage and congestion dis-utility in each city, in turn, are determined jointly with the population distribution across cities in the spatial equilibrium. The inflows of workers affect the local welfare in both directions. Migrants drive down the local welfare by depressing the wage rate and raising congestion dis-utility; at the same time, they also contribute to the local welfare by reducing the price level in the local goods market, as they dampen the production costs and induce more firm entry. Moreover, the local impacts in each city further depend on the migration flows into and out of their neighboring cities

through inter-city trade. In the end, the local impacts of migration remain a question to be answered by the data.

We quantify the model to 279 prefecture-level cities in China in two steps. In the first step, we digitize the road, railway, and waterway atlas from the *Sino Map Press* and estimate the geographic frictions in moving goods and people across the country following Allen and Arkolakis (2014). Conditional on the estimated geographic costs, we then jointly estimate the other parameters using the Simulated Methods of Moments (SMM) in the second step. The parameters governing the migration frictions and congestion disutilities are identified by the data moments related to the bilateral migration flow matrix between 2000 and 2005, such as the city-specific inflow, outflow, and stay rates. The parameters related to the firm entry costs and internal trade costs are identified by the variations in the number of firms across cities and the volume of inter-city trade. We bootstrap the standard errors and show that the data moments can tightly pin down the parameters in our model. In the end, our estimation successfully captures both the targeted moments and several untargeted features in the data as well.

Our main finding is that the local impacts of migration are positive, and all the destination cities benefit from the population inflow. This result comes from the comparative statics between two spatial equilibrium in the year 2000 and 2005. Our benchmark estimation captured the condition of the Chinese economy in 2005 and the bilateral migration flows between 2000 and 2005. To counterfactually simulate the spatial equilibrium in 2000, we increase the migration frictions to revert the bilateral migration flows and thus exactly recover the population distribution in the year 2000.[1] In both the data and our simulation, individuals migrate from the small and inland cities to the large and coastal cities, leading to a more concentrated population distribution across space. Over the five-year span, 40 out of 279 cities received net population inflow that is equal to 9.3 percent of their initial population. The influx of migrants has led to a 4.7 percent increase in real income among

[1]We have assumed that the reduction of migration frictions solely drives the migration flow between 2000 and 2005. In reality, many other factors, such as local demand shocks, productivity booms, and international trade liberalization would also trigger the cross-city migration within China. This conceptual distinction does not affect the welfare analysis since the welfare outcome only depends on the actual flow of migration but not the underlying forces that drive it. However, the distinction does matter for understanding the effects of further migration. As we will show in Section 4, the reduction in migration frictions draws people into existing large cities while international trade liberalization draws workers into coastal cities.

the 40 cities. At the same time, migrants also result in a higher congestion disutility that partially offset the gain in the real wage; the welfare in these cities increases by 4.0 percent after considering congestion disutility, suggesting the additional congestion costs account for $1 - 4.0/4.7 \approx 15$ percent of the gain in the real wage.

The key mechanism behind the positive local impacts is firm entry and exit. The number of firms in each city is endogenously determined by the free entry and exit condition and is responsive to migration flows in our model. The inflows of workers 1) lower the wage rates and at the same time, 2) increase the demand in the local goods markets. The two forces both work to increase the expected profits of the potential entrants and lead to more firms and varieties in the local market, which is a mechanism commonly found in the trade models following Krugman (1980). Residents in the destination cities, in turn, benefit from more varieties at lower prices. In our analysis, the extensive margin of firm entry quantitatively dominates all the negative impacts, leading to the positive local impacts. The mechanism of firm entry is also the fundamental difference between our model and the model in Tombe and Zhu (2019). Using an Eaton-Kortum model without the extensive margin, Tombe and Zhu (2019) find a *negative* relationship between population inflow and real wage, the opposite of our finding. To compare our results, we simulate another counter-factual in which we exogenously fix the number of firms in each city. The local impacts of migration turn *negative* and the population inflows instead *lower* the real wage and welfare in the destination cities in the "no entry/exit" simulation, indicating that the extensive margin of firm entry is the driving force behind our differences. To empirically examine the relationship between the inflow rate, the per capita GDP, and the number of firms in the data, we construct a model-based instrument variable for the city-specific inflow rates using counter-factual simulations following Allen et al. (2014). We find that consistent with our quantitative results, cities that received net population inflow indeed enjoy a higher per capita GDP and the number of firms in the data.

The local impacts of migration depend on the transportation network as well. We are the first to bring the comprehensive Chinese transportation networks — the road, railroad, and water transportation — into the analysis following the methods in Allen and Arkolakis (2014). These networks are essential to the local welfare in many cities, as the wage and

price level depend on inter-city trade through the transportation networks. In the baseline estimation, the top beneficiaries of the internal migration are not the largest cities such as Beijing and Shanghai, but their neighboring cities. The neighboring cities enjoy the productivity boom in the large cities through inter-city trade, and at the same time, are spared from the surge of congestion disutility in the large cities. Such insights can only be obtained when the transportation networks are incorporated into the analysis. Moreover, the estimated geographic costs matrix among the 279 Chinese cities can be easily adapted to the other studies on the Chinese economy.

After evaluating the local impacts of the existing migration flows, we focus on the largest cities and evaluate the implications of the city-specific migration barriers. We first show that the most populous cities indeed enact additional migration barriers as compared to the national average. The barrier into Shanghai is 24 percent higher than the national average, followed by Guangzhou (13 percent) and Beijing (8 percent). The migration barriers result in a 0.7 to 3.9 percent loss in the national welfare. In the case of Beijing and Guangzhou, removing the additional barriers also lead to 9.7 to 18.6 percent gain in the local welfare. However, in Shanghai, the largest city in China, completely removing the barriers might not be desirable. Without the barriers, Shanghai will attract an additional 6 million migrants into the city, and the resulting surge in the congestion disutility dominates the gain in the real wage, leading to a 8.3 percent loss in the local welfare. The diverging impacts of migration barriers at the national and the local level highlight the political difficulty of further reforms on labor mobility. We also show that the largest cities in China are under-populated due to the restrictive migration policy, and relaxing those policies could benefit both the national and the local welfare, which echos the findings in Au and Henderson (2006).

Lastly, we show that the inter-city migration amplifies the welfare gains from international trade. We counter-factually simulate the model with a 10-percent reduction in the international trade barriers. The trade liberalization leads to a 12.7 percent increase in the aggregate income and induces 2.7 percent of the population to migrate, mainly from inland to coastal cities. The gains from trade are 55 percent higher than that without internal migration. Migration amplifies gains from trade through local wage rates. Without migration, the increased labor demand in the coastal cities following the trade liberalization quickly

pushes up the local wage rates, which throttles firm growth in the local markets. With migration, the inflows of workers dampen the wage rates in the coastal cities, which enable further growth of the local firms and lead to higher gains from trade at the aggregate level. Our results add to the recent work on the 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.

Our paper is closely related to the literature on the internal migration in China. Tombe and Zhu (2019) study the impact of migration on aggregate productivity. Fan (2015) studies the effects of international trade on the skill premium and inter-city migration. Our paper instead focuses on the local impacts of migration in the destination cities. To better answer the question of the local impacts, we deviate from the existing work in the modeling choices. Both Tombe and Zhu (2019) and Fan (2015) build on an Eaton-Kortum trade model which does not allow for firm entry and exit; we instead start with a Melitz framework that allows for such a channel. The extensive margin of firm entry delivers a strikingly different quantitative result as compared to the literature. With endogenous firm entry and exit, all the destination cities benefit from the inflows of workers. Once we shut down the extensive margin, our results revert back to those in Tombe and Zhu (2019) that the destination regions suffer from immigration.[2] To this end, our paper highlights the overlooked extensive margin in studying the local welfare impacts of migration. Our second major difference from the literature is the introduction of geography. Tombe and Zhu (2019) focus on the province-level analysis and thus do not introduce the transportation networks into their model. Fan (2015) uses city-to-city distance as a proxy to measure the migration costs. We show that the local impacts of migration also depend on the relative location of the cities on the detailed transportation networks.

Our paper is also related to the broader literature on trade and migration, such as Artu et al. (2010), di Giovanni et al. (2015), Fajgelbaum et al. (2018), Caliendo et al. (forthcoming), and Caliendo et al. (2018). We build on di Giovanni et al. (2015) and introduce endogenous

[2]Fan (2015) studies the immigration as a result of trade liberalization, not the reduction in migration costs as in Tombe and Zhu (2019) and our paper. Although in principle the model of Fan (2015) can also be used to study the local impacts of migration, Fan (2015) did not discuss the local impacts of migration in his paper. We conjecture that the model in Fan (2015) should deliver similar results to Tombe and Zhu (2019) as they are both based on the Eaton-Kortum model without firm entry/exit.

migration decisions similar to Artu et al. (2010) and Tombe and Zhu (2019). Comparing to di Giovanni et al. (2015), our framework is more suitable to study the endogenous response of migration to the small changes in migration frictions. On the other hand, we omit the effects of remittances in our model due to data restrictions. Caliendo et al. (forthcoming) also study the role of frictions in labor and goods mobility in a dynamic setting. They introduce a methodology to carry out counter-factual studies without estimating the level of labor mobility frictions. Different from their work, we model and estimate the migration frictions directly. We do so because we study the variations of the frictions across cities and the resulting welfare impacts, which depend on the *level* of migration frictions. Despite the static nature of our model, the computational load of solving and simulating the model is heavy, similar to most models following Melitz (2003) that allow for firm entry. We overcome the computational difficulty by implementing a new iterative Particle Swarm Optimization (PSO) algorithm. Our algorithm utilizes the solutions within a neighborhood to speed up the computation, and at the same time, avoids local minima by iteration. The algorithm can be easily implemented in parallel, which allows us to estimate our model structurally.

Lastly, our work is also broadly related to the literature on the Chinese economy (Chow, 1993; Brandt et al., 2008; Hsieh and Klenow, 2009; Song et al., 2011). We argue that despite the potential negative impacts on congestion, the residents in the large cities still benefit from the influx of migrant workers, and further reductions in the migration barriers are desirable. We also highlight the interaction between labor mobility and the export-led growth that is being pursued by many local governments. We show that lower frictions in the labor mobility will enable the coastal cities to better benefit from international trade, and lead to gains from both trade and migration.

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 and Section 5 the extensions. Section 6 concludes.

2 The Model

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 similar to Tombe and Zhu (2019).

The economy contains a mass $\bar{L} > 0$ of individual workers and $J > 1$ geographically segmented cities, indexed by $j = 1, 2...J$. Individuals can migrate between cities subject to frictions specified later. There are two production sectors in each city j, namely, the tradable (T) and the non-tradable sectors (N). Firms in the tradable sector can sell to the other cities subject to fixed and variable trade costs, while the firms in the non-tradable sector only serve the local market.

2.1 The Utility Function

Individuals in city j obtain utility from the consumption of a CES aggregate of intermediate goods produced in both sectors according to:

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

where ε represents the elasticity of substitution between the varieties, $y(k)$ is the quantity of variety k, and α captures the expenditure share on the non-tradable sector.

Ω_j^s denotes the set of varieties available for purchase in city j, sector s. It reflects the idea of "love of varieties" common in the CES utility function: everything else being equal, a larger Ω_j^s set is welfare-improving.[3] Ω_j^s is endogenously determined in the model. It depends directly on the firm entry and exit decisions in city j, and indirectly on the number of firms that choose to sell to city j from all the other cities in the case of the tradable sector. The entry, exit, and "exporting" decisions made by the firms are all dependent on the endogenous population distribution and migration patterns, which in turn, rely on the fundamental forces

[3]A larger set of Ω_j^s improves welfare by lowering the ideal price indices, which benefits every consumer regardless of his income in the model. However, it is reasonable to infer that the richer consumers might favor a larger set of varieties more than the poor in reality. In this sense, the quantitative results in the paper are thus more applicable to the incumbents with relatively higher income.

in the model: the underlying migration policy and the transportation networks that we will specify later. In general, a larger market size, potentially as a result of immigration, is able to support more firms and varieties in a city, which is thus welfare-improving given the utility function specification. In the quantitative exercise, we show that the "love of variety" effects are crucial to the local impacts of migration.

Utility is also affected by the congestion disutility from living in city j as denoted by $C(L_j)$:

$$C(L_j) = \rho \cdot \left(\frac{L_j}{\bar{L}}\right)^{\phi},$$

where L_j is the population of city j. We restrict $\rho > 0$ and $\phi > 0$ so that the congestion disutility is increasing in city size.[4] In this setup, $\phi = d\log C(L_j)/d\log L_j$ is the congestion disutility elasticity with respect to population. However, ϕ is not the population elasticity of welfare since $C(L_j)$ enters U_j additively as shown in equation (1). Instead, the welfare elasticity is a function of L_j:

$$\frac{d\log U_j}{d\log L_j} = -\phi\frac{C(L_j)}{U_j}.$$

Everything else being equal, a city with higher L_j will also have a higher welfare elasticity in absolute values due to a higher weight of $C(L_j)$ in U_j. This positive relationship between population and the welfare elasticity reflects the fact that congestion (and housing price) is an issue that is particularly pronounced in the large cities.

2.2 The Firms' Problem

The production side follows Melitz (2003): firms with heterogeneous productivity compete in a monopolistic-competitive market, and each firm produces a unique variety. The one-to-one

[4]It is straight-forward to micro-found the congestion disutility in our model by introducing inelastic endowments in each location and a housing market. We abstract from this in the baseline to avoid the additional computational load. However, this exclusion is not costly to our model fit, as the congestion disutility has already captured the idea and prevented cities from ever-expanding. Moreover, the current model can match the city size distribution and the bilateral migration flows reasonably well without having city-specific endowments as shown in Section 3.4.

mapping between the variety and the firm allows us to interchangeably use k to index both the variety and the firm producing it.

In the tradable sector, "exporting" from city j to city i incurs a fixed cost denoted as f_{ij} in the unit of input bundles specified later. Trade is also subject to the standard iceberg trade cost denoted as $\tau_{ij} \geq 1$: to deliver one unit of goods from city j to city i, the firm must produce and ship τ_{ij} units from city j. In both the tradable and the non-tradable sectors, firms also need to pay fixed costs denoted as f_{ii} units of input bundles in order to sell to the local market. Throughout this section, we describe the problem of a firm in the tradable sector unless otherwise specified. The problem of the firms in the non-tradable sector is a special case in which $\tau_{ij} = \infty$ for all $i \neq j$, and we relay the discussion towards the end of the section.

Demand Maximizing the utility function specified in equation (1) yields the demand function faced by firm k located in city j when selling to city i:

$$q_{ij}^T(k) = \frac{X_i^T}{(P_i^T)^{1-\varepsilon}} \left[p_{ij}^T(k)\right]^{-\varepsilon}, \tag{2}$$

where X_i^T is the total expenditure in city i on tradable goods, $p_{ij}^T(k)$ is the price charged by the firm in city i, and P_i^T is the ideal price index that summarizes the price of all the available tradable goods for consumers to purchase in city i:

$$P_i^T = \left[\int_{\Omega_i^T} \left(p_i(k')\right)^{1-\varepsilon} dk'\right]^{\frac{1}{1-\varepsilon}}. \tag{3}$$

The firm will take the aggregate variables X_i^T and P_i^T as given when deciding its price, $p_{ij}^T(k)$. As usual, higher total expenditure and lower firm-level price lead to higher demand. Moreover, due to the CES utility function, higher price level in the market (P_i^T) also increases the demand for firm k when $\varepsilon > 1$ (which is the case in our quantification). This positive correlation reflects the idea that if *the other firms* charge higher prices in market i, the consumers will substitute away by increasing the demand for firm k.

Production The production of variety k in city j, sector s, is linear in the input bundles denoted as $b_j^s(k)$:

$$q_j^s(k) = \frac{1}{a(k)} b_j^s(k). \tag{4}$$

The input bundle is a Cobb-Douglas combination of the local labor and all of the available intermediate goods from sectors N and T:

$$b_j^s(k) = \left[\ell_j^s(k)\right]^{\beta^s} \left[\left(\int_{k' \in \Omega_j^N} y\left(k';k\right)^{\frac{\varepsilon-1}{\varepsilon}} dk' \right)^{\frac{\varepsilon\eta^s}{\varepsilon-1}} \left(\int_{k' \in \Omega_j^T} y\left(k';k\right)^{\frac{\varepsilon-1}{\varepsilon}} dk' \right)^{\frac{\varepsilon(1-\eta^s)}{\varepsilon-1}} \right]^{1-\beta^s}.$$

In the equation above, $\ell_j^s(k)$ is the employment of firm k and $y(k';k)$ is the amount of variety k' used in the production of k. β^s and η^s are the relative weight of labor and intermediate goods in the production function, respectively. In the tradable sector, the relative contributions of labor and intermediate goods from sectors 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.

Firms are heterogeneous in productivity as captured in $a(k)$, the input bundle requirements for producing one unit of output. Firms with higher productivity need fewer input bundles to produce one unit of output. As in a standard Melitz model, firms draw the productivity $(1/a)$, which is equivalent to the inverse of the input bundle requirement, from a Pareto distribution:

$$\Pr\left(\frac{1}{a} < x\right) = 1 - \left(\frac{\mu}{x}\right)^{\theta},$$

where μ is the location parameter and $1/\mu$ defines the maximum of a. θ represents the tail index. The Pareto distribution implies that the cumulative distribution function (CDF) of a takes the form:

$$G(a) = (\mu a)^{\theta}, a \in [0, 1/\mu].$$

In the baseline model, we abstract away from the agglomeration forces and city-specific

productivity for simplicity. Our main results are robust to these elements as we will show later in Section 5.

Entry and Exit A potential firm first needs to pay f_e units of input bundles to enter sector s in city j. After paying the entry cost, the firm will then draw its productivity from $G(a)$. The firm can then choose whether to continue to produce or to exit depending on its draw of productivity. The outside option from exiting is normalized to be zero. There exists infinitely many potential firms that can choose to enter any city and sector.

2.3 The Solution to the Firms' Problem

We solve the firms' problem in a fashion of backward induction: we first describe the pricing decisions conditional on the firm selling to market i; we then outline the decision of whether or not to sell to market i conditional on firm entry; lastly, we discuss the entry and exit decisions.

Price and Profit in City i If firm k from city j sells to city i, the price, $p_{ij}^T(k)$, is the solution to the following profit maximization problem:

$$\pi_{ij}^T(a) \equiv \max_{p_{ij}^T(k)} p_{ij}^T(k)\, q_{ij}^T(k) - a(k)\, q_{ij}^T(k)\, \tau_{ij} c_j^T,$$

subject to the demand function specified in equation (2). c_j^T is the cost of an input bundle in the tradable sector at city j, which itself is the solution of a cost minimization problem:

$$c_j^T = (w_j)^{\beta^T} \left[\left(P_j^N\right)^{\eta^T} \left(P_j^T\right)^{1-\eta^T} \right]^{1-\beta^T},$$

where w_j is the wage rate in city j, and P_j^N and P_j^T are the ideal price indices in the N and T sectors. It is straightforward to see that the solution of the pricing problem is a constant markup of $\varepsilon/(\varepsilon-1)$ over the marginal cost of production, $c_j^T a(k)$, adjusted by the iceberg

trade cost:

$$p_{ij}^T(k) = \frac{\varepsilon}{\varepsilon - 1} \tau_{ij} c_j^T a(k).$$

A more productive firm with a lower $a(k)$ is able to charge a lower price, and thus capture a larger market share in city i. The variable profit is also higher for firms with lower $a(k)$ as π_{ij}^T is proportional to $(a(k))^{1-\varepsilon}$:

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

At the city-level, if the targeted market size is larger (X_i^T higher), or the other firms in the market are relatively unproductive (so they charge higher prices, leading to a higher P_i^T), the variable profit for firm k will be higher since the demand for the firm's product is larger as indicated by the demand function in equation (2).

"Exporting" and the Total Profit Conditional on the solution of the pricing problem, a firm with input bundle requirement $a(k)$ in city j will serve city i if and only if the variable profit can cover the fixed cost of trade, f_{ij}:

$$\pi_{ij}^T(a) \geq f_{ij} c_j^T,$$

Moreover, the inequality also implies a cutoff rule: the firm in city j will sell to city i if and only if its $a(k)$ is below a_{ij}^T:

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

Lower transportation costs (τ_{ij}) or operating barriers (f_{ij}), an expansion of the market size (X_i^T), or a reduction in the unit production cost (c_j^T) all lead to a higher a_{ij}^T, allowing more firms to sell from city j to i. A firm in city j compares its input bundle requirement to all the cutoffs $a_{ij}^T, i = 1, 2, \cdots, J$ to determine the market(s) to sell to.

The sales decisions at this stage imply that the total profit of the firm with unit cost

$a(k)$, net of the entry costs, is the summation over all the potential markets i:

$$\Pi_j(a(k)) = \sum_{i=1}^{J} \mathbf{1}\left(a\left(k\right) < a_{ij}^T\right)\left(\pi_{ij}^T(a) - c_j^T f_{ij}\right),$$

where $\mathbf{1}\left(a\left(k\right) < a_{ij}^T\right)$ is an indicator function that equals to 1 if the draw is low enough to serve city i, and 0 otherwise. More productive firms will be able to sell to more markets and earn a higher total profit. If a firm is not productive to gain profit in any market, it chooses to exit immediately.

The Entry Decision At this final stage, we are able to characterize the entry decision of the potential firms. Prior to paying the entry cost f_e and draw $a(k)$, the *expected profit* of a potential entrant in city j is:

$$\overline{\Pi}_j^T \equiv E\left[\Pi_j(a(k))\right] = \int_0^{1/\mu} \Pi_j(a) dG(a).$$

The expectation is taken over the distribution of $a(k)$ as characterized by $G(a)$. In the equilibrium, the expected profit in city j must be equal with the entry cost due to zero outside option and infinitely-many potential entrants:

$$\overline{\Pi}_j^T = f_e c_j^T.$$

Finally, the ideal price index of the tradable sector in city j is the aggregation over all the varieties sourced from all the cities (including itself) as indexed by i:

$$P_j^T = \left[\sum_{i=1}^{J}\left(\frac{\varepsilon}{\varepsilon-1}\tau_{ji}c_i^T\right)^{1-\varepsilon} I_i^T \int_0^{a_{ji}^T} a^{1-\varepsilon} dG\left(a\right)\right]^{\frac{1}{1-\varepsilon}},$$

where I_i^T is the measure of firms that entered the tradable sector in city i, and a_{ji}^T is the cutoff below which the firm in city i will sell to city j. The "love of variety" effect is reflected in the above expression: if more firms are able to sell to market j through either a higher number of entrants (I_i^T) or a higher cutoff (a_{ji}^T), then the ideal price index in city j will be lower as $\varepsilon > 1$.

The N Sector The problem of the firms in the non-tradable sector is a special case of the tradable sectors in which $\tau_{ij} = \infty$ for all $i \neq j$. The firm charges the price:

$$p_j^N(k) = \frac{\varepsilon}{\varepsilon - 1} c_j^N a(k),$$

in the local market, where c_j^N is similar to c_j^T with β^N and η^N instead. The firm earns the following variable profit:

$$\pi_j^N(a) = \frac{X_j^N}{\varepsilon \left(P_j^N\right)^{1-\varepsilon}} \left(\frac{\varepsilon}{\varepsilon - 1} c_j^N a(k)\right)^{1-\varepsilon},$$

and it only operates if its draw of $a(k)$ is below the cutoff a_j^N:

$$a_j^N = \frac{\varepsilon - 1}{\varepsilon} \frac{P_j^N}{c_j^N} \left(\frac{X_j^N}{\varepsilon c_j^N f_{jj}}\right)^{\frac{1}{\varepsilon - 1}}.$$

The entry decision is characterized as:

$$\overline{\Pi}_j^N \equiv E\left[\mathbf{1}\left(a(k) < a_j^N\right)\left(\pi_j^N(a) - c_j^N f_{jj}\right)\right] = f_e c_j^N.$$

Lastly, the price index is based on the varieties produced locally:

$$P_j^N = \left[\left(\frac{\varepsilon}{\varepsilon - 1} c_j^N\right)^{1-\varepsilon} I_j^N \int_0^{a_j^N} a^{1-\varepsilon} dG(a)\right]^{\frac{1}{1-\varepsilon}},$$

where I_j^N is the mass of firms that entered the non-tradable sector in city j.

2.4 Migration Decision

The migration decision depends on three components in our model: the indirect utility, the idiosyncratic preference, and the bilateral migration friction. The indirect utility of living in city i, which we denote as U_i, comes from the equilibrium of the model conditional on population distribution. The indirect utility depends on the real wage rate and the congestion

disutility:

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

In addition to the indirect utility, each worker also draws an idiosyncratic preference shock toward each city $\{\iota_i\}_{i=1}^{J}$, where ι_i is $i.i.d$ across locations and individuals. 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.

Lastly, moving from city j to i incurs origin-destination specific costs in the unit of utility, which we denote as λ_{ij}. The costs of migration enclose not only the financial costs of moving but also the various policy barriers that deter migration such as Hukou, working permits, and other bureaucratic red tape. Combining the three elements discussed above, 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,

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

It is straightforward to show that conditional on U_i, the fraction of 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 similar to the one used in Tombe and Zhu (2019) and 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 indirect utility in the destination city and negatively related to the bilateral frictions, which depend on the distance and policy barriers in our context. Both of these assumptions are strongly supported by the data in the context of international

migration.

2.5 Equilibrium

Definition: Given all the parameters of the model, the equilibrium contains a series of prices $\{w_j, p_{ij}^T(k), p_j^N(k)\}_{j=1}^J$, and a sequence of quantities $\{I_j^T, I_j^N, L_j, q_{ij}^T(k), q_j^N(k)\}$ such that the following conditions hold:

(a) Individuals maximize their utility by choosing locations and consumption bundles from both sectors.

(b) Each intermediate goods producer maximizes its profits by choosing its prices and the markets to sell to.

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

(d) Goods market clearing:

$$\begin{aligned} X_j^N &= \alpha w_j L_j + \left(1-\beta^N\right)\eta^N X_j^N + \left(1-\beta^T\right)\eta^T X_j^T, \\ X_j^T &= (1-\alpha)\, w_j L_j + \left(1-\beta^N\right)\left(1-\eta^N\right) X_j^N + \left(1-\beta^T\right)\left(1-\eta^T\right) X_j^T. \end{aligned}$$

(e) Labor market clearing:

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

3 Quantification

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 they cannot move between China and the ROW.[5] In the rest of this section, we first outline the estimation of the geographical struc-

[5]The United Nations estimated that the emigration rate from China during this period is between 0.0047% and 0.0093% (IMO, 2006). Comparing to the magnitude of internal migration at around 12% during the same period as estimated in this paper, the migration between China and the ROW is negligible. For simplicity, we assume away the international migration.

ture, both within China and between China and the ROW, and we then describe the empirical issues in estimating the parameters related to the population distribution and bilateral migration flows. 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 The 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 in both the *Chinese City Statistical Yearbooks* and the *One-Percent Population Survey* carried out in 2005 (thereafter the 2005 Micro Survey). Our sample, as shown in Figure 1, 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 in our sample; those missing are mainly the cities in Tibet, Xinjiang, 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 measure the shortest distance between the city pairs using different transportation modes. Third, we combine the first two steps and structurally estimate the parameters that govern the relative costs of using different transportation modes, and then arrive at 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 the good. The traders choose a particular transportation mode to minimize the costs incurred from shipping. 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:

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

where $d_m(i,j)$ is the distance from city j to i using the 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 specification above allows us to express 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 as:

$$\frac{\exp(-a_m d_m(i,j) - b_m)}{\sum_{n=1}^{M}(\exp(-a_n d_n(i,j) - b_n))}, \tag{5}$$

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 use three modes of transportation: road, water, and railway. For each mode, we identify the location of the existing infrastructure using the high-resolution transportation maps from the *2005 China Maps* published by Sino Map Press. Each scanned 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 the normalized road network, 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 reflect the differences in speed limits under Chinese law.[6] We normalize the pixels with navigable waterways, including open seas, to a cost of 1, and all the other pixels with a cost of 10 following Allen and Arkolakis (2014). To construct the raster for normalized railroad cost, we assign all pixels with rail road access a cost of 1, and all the other pixels a cost of 10. Lastly, we 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, $d_m(i,j)$.

[6]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.

Given the mode-specific distance matrix, we next estimate the cost parameters $\{a_m, b_m\}$ in equation 5. Following Allen and Arkolakis (2014), we estimate these parameters by matching the fraction of trade volume in each city that goes through the transportation mode m in the data. The city-mode-specific trade volume comes from two data sources. The *China City Statistical Yearbook 2005* reports the quantity shipped in metric tons in each city under the transportation mode m.[7] To infer the trade flows in monetary values, we also need to estimate the value per ton of goods that goes through mode m. To do this, we turn to the transaction-level custom dataset in China which reports the value, quantity, and mode of transportation for the universe of Chinese imports and exports. In 2005, the results from 22.82 million custom transactions indicated 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 have arrived at the 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 $V_m(i)$, 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 fraction of trade volume under transportation mode m in city i, $s_m(i)$, can thus be expressed as:

$$s_m(i) = \frac{V_m(i)}{\sum_{m'=1}^{M} V_{m'}(i)}. \tag{6}$$

The variable $s_m(i)$ is strictly between 0 and 1 for all the cities. The variation of $s_m(i)$ across cities reflects the fact that cities utilize the transportation networks differently. For example, Beijing, a non-coastal city without any direct connection to a waterway, relies less on water transportation than Shanghai, and thus feature a smaller s_{water}.[8]

[7]The total quantity shipped in city i by mode m in the data includes goods shipped from city i to all other cities and the those delivered to city i from all other cities as well.

[8]The volume of trade in the data, $V_m(i)$, also depends on city-specific components such as population, GDP, and productivity. One might be concerned that this will bias the estimation of a_m and b_m. However, to the extent that the city-specific factors enter $V_m(i)$ multiplicatively, which is the case in most trade models, they would show up in both the numerator and the denominator, and thus be canceled out in computing

We estimate $\{a_m, b_m\}$ using a non-linear least squares 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 $\{a_m, b_m\}$ 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 to our model (75.4 percent). We are also able to capture the relative weight of rail and river transportation with error margins at around one percentage point.

We follow the estimations in Allen and Arkolakis (2014) and set θ_T to 17.65.[9] Given $\{a_m, b_m\}$ and θ_T, the discrete choice framework implies that the average geographic costs from city j to i can be obtained as follows:

$$T(i,j) = \frac{1}{\theta_T}\Gamma\left(\frac{1}{\theta_T}\right)\left(\sum_m \exp(-a_m d_m(i,j) - b_m)\right)^{-\frac{1}{\theta_T}}, \tag{7}$$

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

Of the three modes of transportation, the estimated T matrix mostly depends on the road network. The significance of the road network is because all the cities in our sample have direct access to the national road system, and the vast majority of intra-China trade also goes through the road network. As seen in Figure 2, the trade costs increase with distance as measured in $d_m(i,j)$ regardless of the transportation mode. However, in contrast to the road transportation, geographical costs vary significantly between city pairs with similar rail or waterway distances. Traveling by river or coastal sea has the largest variation, mainly because many Chinese cities do not have direct access to any navigable waterway.

Empirical work often uses physical distances between cities as proxies for the transportation costs, implicitly assuming that the city pairs with similar physical distance also share

$s_m(i)$. In other words, as long as the city-level factors enter $V_m(i)$ uniformly across modes of transportation, our estimates of a_m and b_m will not be affected.

[9]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 innocuous. Equation (7) 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.

similar difficulties in transportation. Indeed, the geographic costs increase with physical distance. as seen in the last panel of Figure 2. However, conditional on a given physical distance, the variations in geographic costs are large and increasing with physical distance. The variations in our estimated geographic costs are rooted in the placement of transportation networks. For example, the geographical costs for city pairs 1,000km apart can be as low as 1.15 if the pair is well-connected through the transportation networks, or be as high as 1.37 if the pair is poorly-connected. For city pairs 3,000km apart, the geographic costs can range between 1.54 and 2.0. The large variations indicate 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 collapse 148 trading partners of China into 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 distance database (which we discuss later) are in the sample. We estimate the geographical distances following a similar strategy as described above 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 seaports in the world are widely available, much less can be obtained for the other two modes of transportation. This assumption is also innocuous: records from Chinese customs indicate that on average, over 80 percent of international trade measured in value and over 90 percent measured in weight goes by sea between the year 2000 and 2005.

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 coastal city i in China, which is denoted as r_{ik}.[10] The distance between

[10]For countries facing multiple oceans or with long coastlines, 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 route, going through the Suez Canal, the Panama Canal,

ROW and the coastal city i is then:

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

ξ converts nautical miles, which is the unit of r_{ik}, to the units used in $d_{\text{sea}}(\cdot)$ for waterborne transportation in China.[11] 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 denoted as Λ_k. Lastly, we use equation (7) again with the $d_{\text{sea}}(i, \text{ROW})$, assuming the distances in the other two modes to be infinity, to compute the T_{ij} between any coastal city in China to the ROW.

For an 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 the 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.

3.2 Population and Migration

In addition to the geography data, we also need the initial population distribution in a given year, and the bilateral migration flows between the initial year and a later year to discipline the model. For the initial population distribution, we use the year 2000 distribution over the 279 cities from the 2000 population census. For the between-city migration flow, we rely on 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 a problematic estimate due to the changes in city boundaries.

or the Strait of Gibraltar.

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

The primary challenge in our data construction is that the official definitions and boundaries of cities changed significantly between 2000 and 2005. In our sample of 279 cities, the geographic boundaries of 12 cities are re-drawn, rendering them incomparable between the two data sources. Moreover, 49 new prefecture-cities are established between 2000 and 2005, and therefore they did not exist as prefecture-level administrations in the 2000 census. To solve these problems, we construct a geographically-consistent dataset of city populations between 2000 and 2005 based on the city boundary in 2005 (the "2005-cities" hereafter). The official records from the central and provincial governments contain information on how sub-city administrative units (counties, Xian) are grouped into new cities or how they are re-assigned between the existing cities. We use these records to map counties in 2000 to their respective cities in 2005. We then reconstruct the population of 2005-cities based on this county-city mapping and the population of each county in the 2000 population census. The implicit assumption of our re-construction exercise is that the boundaries of the counties remained unchanged between the year 2000 and 2005, which, to our best knowledge, is correct. The resulting data set is the first geographically-consistent population panel data at the city level.

To incorporate ROW in our population data, we add the total population of the 148 trading partners of China as the raw population of the ROW.[12] We allow for potential differences in the total factor productivity (TFP) between the ROW and China by introducing a parameter to measure the relative efficiency between the Chinese and ROW workers. In the end, the initial population used in our model 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 population of the Chinese cities in 2000 that we have constructed above, ℓ_{ROW} is the total population of the 148 trading partners, and A is the relative

[12]Data source: *World Development Indicators, 2000.*

TFP that we estimate later.

3.3 Quantifying the Structural Parameters

Our parameter space contains the following structural parameters:

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

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

3.3.1 Calibration

ε is the elasticity of substitution between the varieties, and 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 $\theta/(\varepsilon-1)$. We follow di Giovanni and Levchenko (2012) by setting ε to 6. We set θ to 5.38 so that the tail index of the employment distribution is 1.076, the value computed from the *Annual Surveys of Manufacturing Firms* in China.

β_N and β_T reflect the share of labor in production, 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 ratio 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 the non-tradable and tradable sectors, and we also calibrate them using the same 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, the intermediate goods from the non-tradable sectors play a larger role in the production of the other non-tradable goods in the Chinese data as well. In contrast to the U.S. data, the non-tradable goods are overall less important in both sectors, probably because many service industries, such as finance and consulting, are relatively less developed in China. α governs the expenditure share on the non-tradable goods. We set it to be 0.61, the share of the total consumption of the non-tradable goods, which is computed from the final use table in the same year. The housing sector is counted as non-tradable in the IO table.

We calibrate the f_{ij} matrix following the strategy outlined in di Giovanni and Levchenko (2012). Everything else being equal, higher fixed operating costs lead to a lower number of firms in equilibrium, and thus a smaller fraction of individuals working as entrepreneurs in that city. di Giovanni and Levchenko (2012) used the difficulties of starting new businesses across countries from the *Doing Business* data set to calibrate the fixed operating costs. As similar measure does not exist at the city-level in China, we instead turn to the 2005 micro survey and approximate $1/f_{ii}$ by using the fraction of entrepreneurs in each city among all working population.[13] To discipline the off-diagonal elements, f_{ij}, we first set them as the sum of the two diagonal elements f_{ii} and f_{jj}. At this stage the f_{ij} matrix is not in the unit of the input bundle as in the model, and to convert it to the correct unit, we scale the entire matrix with a factor ζ. We set ζ to ensure interior solutions in all of the counter-factual simulations. We summarize all the calibrated parameters and their corresponding targets in Table 2.[14]

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

[13] We define entrepreneurs as those reported as "employers" (Item 2 in Question R23); this definition excludes the self-employed.

[14] Interior solution means that $a_{ij} \leq 1/\mu$, 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.

to the geographic costs matrix up to a scale parameter $\bar{\tau}$. This assumption is based on the widely documented fact in the trade literature that the physical distance reduces bilateral trade. The second line specifies the trade costs between the Chinese cities and the ROW. As national borders usually introduce significant costs to international trade, we allow for an additional international trade barrier, τ_{row}, to capture the border effect.[15] The above simplification reduces 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 contain two parts. The first part, $\bar{\lambda} \cdot T_{ij}$, is symmetric between i and j and proportional to the geographic cost T_{ij}. All else being equal, it is easier to move to nearby cities because of the ease of travel and the 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. The caveat of directly using the geographic cost matrix (T) is that the matrix is estimated from the traffic volumes of goods instead of passengers, and moreover, we have omitted air transportation altogether. However, directly using the T matrix is still innocuous for two reasons. First, the relative importance of the road, railway, and waterborne transportation for passengers are roughly the same as for goods. For example, in 2005, 91.5 percent of passenger transportation goes by road, followed by 6.7 percent by railroad and 1 percent by waterway.[16] The ranking is the same as the goods transportation as seen in Table 1 and the magnitude is similar as well. 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 of λ_{ij} is the *destination-specific* migration cost, δ_i. A large part of

[15] See McCallum (1995) and Anderson and van Wincoop (2003) for examples.

[16] The data source for passenger traffic is the same as goods traffic: *China City Statistical Yearbooks*.

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

In the end, the vector of the 12 parameters to be estimated by SMM is

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

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

$$\widehat{\Theta} = \text{argmin}_{\Theta}[S - \widehat{S}(\Theta)]\widehat{W}[S - \widehat{S}(\Theta)]'. \tag{8}$$

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.[17] 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 of our algorithm in the appendix.

The S vector contains 20 data moments that are important in disciplining the Θ vector. We first discuss the moments that do not rely on the bilateral migration data, and then turn to the migration-based moments.

Non-Population Based Moments 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 sales by

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

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 come from 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, using "legal entities" (Fa Ren) as definition of firms. This statistic identifies f_e, which captures the fixed costs of firm entry in the unit of input bundle. 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 infrastructure and institution 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 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 between China and the ROW in 2005. The first moment, 59.4 percent, comes from the National Bureau of Statistics. The second moment is based on the data from the WDI, which states that the 148 trading partners combined are around 21.32 times larger than China in GDP. These two moments help to identify the international trade barrier, τ_{row}, and the relative productivity, A.

Population-Based Moments The other moments in the S vector depend on the bilateral migration flows between 2000 and 2005 that we have constructed in the previous section. We denote the migration matrix M in the data as follows:

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

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 $\sum_{i=1}^{J} \ell_{ij}$ is the population of city j in

the year 2000, while $\sum_{i=1}^{J} \ell_{ji}$ is the population of city j in the year 2005. We denote the population distribution vector in these two years as L_{2000} and L_{2005}, respectively. We focus on the following 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:

$$\text{Aggregate Stay Rate} = \frac{\sum_{i=1}^{J} \ell_{ii}}{\sum_{i=1}^{J} \sum_{j=1}^{J} \ell_{ij}}.$$

As higher costs of migration deter overall migration and lead to a higher aggregate stay rate, this moment pins down the overall costs of migration, $\bar{\lambda}$. At the aggregate level, we also target the Pareto tail index of the city size distribution in 2005 to capture the overall shape of the population distribution.

In addition to the aggregate stay rate, we also target the stay rates by groups of cities. The stay rate in city i, defined as:

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

measures the propensities to emigrate out of the city. The cities with low stay rates are the popular *origins* of the migrants. The 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 out to the other cities in 2005, but the emigration rate for some smaller cities can be as high as 49 percent. The variations in stay rates reflect the basic pattern of bilateral migration in China: people moving from the small and inland cities to the large and coastal ones. To capture this salient feature, we rank the 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 the city-specific stay rates, both across the entire nation and within each of the four groups of cities to capture the within-group variation in the stay rates.

The last group of moments are based on the inflow rates into city i, which is:

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

Different from the stay rate, the inflow rate reflects the popularity of the city as a *destination* among migrants. We use the inflow rates in two ways. First, we target the correlation between the logarithm of the city population in 2000 and the city-specific inflow rate. This correlation captures the feature in the data that the cities with higher initial population usually have higher inflow rates as well. Secondly, we also target the inflow rates of the four cities on which we have imposed the 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 to identify the parameters governing the 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 population of these cities. All the data moments are summarized in Table 3. All of the estimated parameters, along with the bootstrapped standard errors, are reported in Table 4.[18]

The Jacobian Matrix and Identification We compute the elasticity of all the targeted moments with respect to the parameters around the baseline estimation to highlight the source of identification. The elasticity is approximated by 0.1 percent local variations around the baseline. Table 5 reports the results.

The parameters affect the moment conditions in directions predicted by the model. Higher migration frictions ($\bar{\lambda}$) lead to higher stay rates at both the national level and across all groups of cities. Higher congestion cost (ρ) throttles migration into the large cities as well. Higher frictions of firm entry (f_e), in addition to reducing the number of firms, also discourage population inflow toward the large cities and aggregate migration, because the benefit of migration depends positively on new firm entry. City-specific entry barriers barely affect any aggregate moments but are instrumental to the inflow rates into the city in ques-

[18] See the appendix for the details of the bootstrapping algorithm.

tion. As expected, the parameters governing the ROW mostly affect the relative size and the openness of China. Interestingly, both the internal and international trade barriers (τ and τ_{row}) affect the population of the coastal cities such as Shanghai, Guangzhou, and Shenzhen with an elasticity between 4.17 to 6.67 through the trade-induced migration. The impact of the trade barriers on Beijing, a non-coastal city, is much smaller with an elasticity between 0.47 and 0.86. Overall, the elasticity matrix highlights the fact that many moments are determined by parameters governing different aspects of the model in general equilibrium, and thus a joint estimation procedure based on simulations is a necessity in disciplining the parameters of the model.

3.4 Model Fit

We evaluate the fitness of our quantification on both the targeted and the untargeted moments. Table 3 that reports the 20 targeted moments also summarizes the model fit in the last column. Overall, our model can match the data moments with relatively small discrepancy. For most of the moments, the relative difference between the model and the data is 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 untargeted moments in the data: the population and GDP distributions in 2005 and the entire bilateral migration flow matrix between 2000 and 2005. These results are summarized in Figure 3.

Our baseline model matches the population distribution in the year 2005 in the data. The model prediction and the data consistently line up along the 45-degree line with a correlation coefficient of 0.838 as shown in Panel (a) of the figure. Similarly, we are also able to match the city-level GDP with a correlation of 0.860, as shown in Panel (b). Most importantly, our calibration strategy captures the bilateral migration flow between all the city-pairs as well: the correlation between the model and the data is 0.660, as seen in Panel (c). We have assumed that the bilateral migration matrix to be proportional to the geographical trade costs, which might seem to be oversimplifying at first glance. However, the model fit validates our assumption, as we can capture the bulk of variations in the data after all. The goodness-of-fit also suggests that the migrants do prefer neighboring cities and cities located

along the main traffic networks. The downside of our fit is that we tend to under-predict when the migration flow in the data is relatively small. The bias is probably because our targeted moments are geared towards the large cities and the major migration flows. Given that we only target a small number of moment conditions, we view these discrepancies as an affordable price to pay for our parsimonious quantification strategy.

4 Quantitative Results

We discuss our results in this section. We first evaluate the local and national impacts of existing migration flows. We then turn our attention to the largest cities and use the counter-factual experiments to understand on the impacts of current policy barriers. Lastly, we study how migration interacts with the internal and international trade liberalization.

4.1 The Local Impacts of Migration

We study the local impacts of inter-city migration between 2000 and 2005 by comparing our baseline model in 2005 to a counter-factual simulation in which the population distribution is numerically identical to that in the year 2000. In the counter-factual simulation, we set the migration cost multiplier $\bar{\lambda}$ to a sufficiently large value so that individuals will not migrate from their initial locations in 2000. This is equivalent to sending all of the migrants in the benchmark model back to their 2000 locations. We compare the results to our benchmark quantification to study the impacts of the intercity migration between 2000 and 2005.[19]

Before we turn to the welfare impacts, we first highlight the key pattern of the intercity population flow: workers migrate from the small to the large cities. The concentration into the large cities shows up both in the data and in our simulation as shown in Figure 4. The population of the largest cities such as Beijing and Shanghai grew by around 15 percent, and the population of Shenzhen almost doubled during the period in both the data and

[19]Our comparative statics exercise does not assume that the population distribution in the year 2000 is an equilibrium distribution. Instead, the comparison between the counter-factual and the benchmark only assumes that the migration frictions in 2005 are lower as compared to those in 2000, and the lowered migration frictions induced *additional* migration between the two years. For this reason, our results shall be interpreted as the implications of the migration *flow* between the two years, not as the implications of existing *stock* of migrants in 2005 as in Tombe and Zhu (2019).

the model. In contrast, the small cities generally lose population. 233 out of the 279 cities experienced emigration in the data; similarly, in our model 239 cities lost population. The concentration of population into the already populous cities is the essential feature of the data that drives most of our results, as we will describe in detail later. Interestingly, the pattern of migration found here – from smaller and less productive locations to larger and more productive ones – is also reflected in the cross-border migration as documented in di Giovanni et al. (2015).

Our key finding is that the local impacts of migration are positive: all the cities that received net population inflow enjoy higher real wage and welfare due to immigration. The first column of Table 6 reports the baseline results. Between 2000 and 2005, 40 large cities received net population inflow, and all of them enjoy a higher real wage. Most of these cities are coastal cities in the east or the provincial capitals in the landlocked provinces as shown in Figure 5. On average, the population in all the "receiving" cities increases by 9.3 percent, and the real wage rate increases by 4.7 percent. The gains from migration are higher in cities with higher inflow rates, as the correlation between the net inflow rate and the changes in real wage is 0.977. The city with the highest inflow rate, Shenzhen (90 percent), also experienced the largest gain in the real wage, 48 percent.

The inflows of workers into the large cities indeed impose additional congestion disutility; however, the cost of congestion is small relative to the gain in the real wage. The surge in population leads to congestion disutility that offsets 15 percent of the gain in the real wage, leaving the receiving cities with $(1 - 0.15) \times 4.7 \approx 4.0$ percent of the gain in welfare as shown in the bottom two panels in Figure 5 and in Table 6. In the largest cities, the surge in congestion disutility indeed offsets a more substantial fraction of the real wage gain. For example, the population of Shanghai increased from 14.5 to 16.8 million in our model, leading to a real wage increase by 8.8 percent. The population growth, at the same time, imposes an additional congestion disutility equal to around 6.0 percent of the pre-migration real wage, wiping out $6.0/8.8 \approx 68.7$ percent of the economic gain. Similarly, congestion disutility offsets the real income growth in Beijing by 29.9 percent, and in Shenzhen by 30.4 percent. Nevertheless, even in the most crowded cities, the residents still stand to benefit from the population inflow after considering congestion.

The mechanism of firm entry and exit drives the positive local impacts of migration. The number of firms operating in each city is endogenously determined in equilibrium and thus responds to population movements. Population inflows reduce nominal wage and enlarge the market size, which in turn support more firms and varieties in equilibrium. A higher number of varieties in the goods market lowers the ideal price index and thus benefits the locals, which is the "love of variety" mechanism from Krugman models. The extensive margin of firm entry dominates the negative impacts of population inflow.

To highlight this point, we carry out another counter-factual simulation in which we fix the number of firms in each city to their levels in the year 2000 and then repeat the benchmark exercises by decreasing $\bar{\lambda}$. The results are reported in the second column of Table 6 and Figure 6. Without firm entry, the vast majority of destination cities see lower real wage, and the inflows of migrants instead *lower* the real income and welfare in the destination cities. The negative local impacts of migration are similar to the findings in Tombe and Zhu (2019) who abstract away from firm entry. The sharp contrast of the results with and without firm entry underscores the importance of this mechanism in understanding the local impacts of migration.[20]

The geographic location of the cities also matters. Figure 7 plots the top 10 winners and losers regarding welfare. Several cities among the top beneficiaries list gain from migration not because they are the prime destinations of the migrants, but because of their advantageous locations. Tianjin, Suzhou, Hangzhou, and Foshan are only several hundred kilometers away from the "star" cities like Beijing, Shanghai, and Shenzhen. These "neighbor" cities are the most direct recipients of the productivity spillovers in the booming "star" cities, and at the same time, they are spared from the surging congestion disutility in the "star" cities. Wuhan and Nanjing are not the next-door neighbors of the largest cities; however, they are among the most strategically placed cities in China. Wuhan sits at the crossroad of the 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,

[20]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.

and it is also on the Beijing-Shanghai railway. Their strategic locations allow them to gain from the productivity booms in the "star" cities similar to the "neighbor" cities.

Cities lose from migration mainly due to two reasons: they either lose population to the large coastal cities in the east, or they are located in remote positions, which in turn deny them easy access to the productivity spillovers from the east. The top "loser" in Figure 7 is Xinzhou, followed by Lvliang and Yulin, which are all located in the remote western frontier of China proper. A combination of the two reasons applies to all of the losing cities to different degrees. However, when interpreting the negative impacts on the origin cities, we have omitted the remittance from migrants to home due to the lack of data at the city level.[21] This implies that the negative impacts in our paper are over-estimated, and in the real world those impacts are partially off-set by the inflow of remittance.

Empirical Tests on the Local Impacts In the previous part, we show that the local impacts of migration are positive, mainly through the mechanism of firm entry and exit. In this part, we provide empirical evidence for this result by studying the relationship between the inflow rate, the per capita GDP, and the number of firms at the city level. The starting point of the estimation is a simple regression to single out the relationship between the inflow rate and the outcome variables in a city:

$$\text{outcome}_i = \beta_0 + \beta_1 \widehat{\text{Ln}\left(\text{Pop.}\right)}_i + \beta_2 X_i + \varepsilon_i, \tag{9}$$

where outcome_i is either the percentage change of per-capita GDP or the percentage change of the number of firms in the city i.[22] $\widehat{\text{Ln}\left(\text{Pop.}\right)}_i$ is the percentage change in population, which is equivalent to the inflow rates as defined in the previous sections, and X_i is a vector of city-

[21] The existing estimates on the magnitude of remittance vary dramatically from around 17 to 19 percent (Taylor et al., 2003; Akay et al., 2014) to 34 percent among the remittance-receiving households (Liang et al., 2013). All of these studies focus on the rural-urban migration, and most of the datasets are aggregated at the province level, which make it hard to map these estimates to our finding directly. Moreover, since only a fraction of migrants remit back to home, and only a fraction of households send out migrants, the overall impact on per-capita income in a prefecture city is even less clear. Nevertheless, as emphasized in di Giovanni et al. (2015), a comprehensive welfare analysis of migration should not ignore the issue of remittances. In the context of China, this calls for more empirical studies on the impacts of remittance in the future.

[22] The percentage change of per-capita GDP comes from the city statistical yearbooks. The changes in the number of firms come from the first and the second economic census. We define the firms as "legal entities" (Fa Ren), the same as in the quantification part.

level controls, which includes the initial population, GDP, and the location on the traffic network measured by the remoteness index that we have estimated. Directly estimating equation 9 using OLS leads to bias in $\hat{\beta}_1$ due to the classic endogeneity issue: unobserved variables buried in ε_i might drive the population change and the outcome variables at the same time; reverse causality is another concern as the population movements might be driven by the changes in economic conditions as well.

To tackle the endogeneity issue, we construct an instrument variable (IV) for the inflow rates in the data. Following the work of Allen et al. (2014), we use a model-based IV. We simulate our model to generate the IV in the following steps. In the first step, we start from a candidate vector of the key parameters $\{\theta^{IV}, \kappa^{IV}, f_e^{IV}, \bar{\lambda}^{IV}, \bar{\tau}^{IV}, \phi^{IV}, \rho^{IV}\}$ and simulate the model to generate a baseline distribution of population.[23] We then decrease the friction of migration to $\bar{\lambda}^{IV'}$ to generate migration flow between the cities, and back-out the simulated inflow rates from this exercise. We use the simulated inflow rate as the instrument variable for the observed inflow rate in the data. The model-based inflow rates meet the exclusion restriction. Unlike the baseline estimation, none of the key parameters in the IV-simulation are based on the actual population flow in the data by design. The simulated population flow is not subject to reverse causality either, as they are only driven by the changes in $\bar{\lambda}^{IV}$. Lastly, the changes in migration friction ($\bar{\lambda}^{IV}$) can only affect the economic outcomes through population movement by the design of the model. Similar to Allen et al. (2014), our model-based IV are functions of the initial population and the bilateral matrix of geographic costs. Since we control for the initial population and location in our estimation, the identifying assumption is that the interaction between the changes in $\bar{\lambda}^{IV}$ and the initial population and the location of the *other cities* are not correlated with the error term, ε_i.

We find that higher inflow rates indeed lead to higher per-capita GDP and a higher number of firms in the destination cities. The first four columns of Table 7 report the OLS estimate. The simple OLS estimates already suggest that cities that received net population inflow enjoy higher growth rates in both per-capita GDP and the number of firms. The next

[23] In the IV simulation, we use $\{\theta^{IV} = 5.3, \kappa^{IV} = 0.003, f_e^{IV} = 2, \bar{\lambda}^{IV} = 0.5, \bar{\tau}^{IV} = 1.5, \phi^{IV} = 3, \rho^{IV} = 150\}$. Note that we purposefully avoid using the baseline estimation, as those parameters are disciplined by the inflow rate in the data. We have experimented with various candidate parameters for the IV simulation, and the results are similar to those presented in Table 7. We later reduce the migration friction, $\bar{\lambda}^{IV'} = 0.02$ to generate the bilateral migration flows.

four columns of the table report the IV estimate. The F-statistics in the first stage of the IV regressions are all in the high range, indicating that the model-based inflow rates are strongly correlated with the data. The point estimates of the IV regressions are also significantly different from the OLS results, suggesting that the OLS estimates are probably biased due to the endogeneity issue. Based on the IV regression, we find that a one-percentage-point increase in the inflow rates increases the growth rate of per-capita GDP between 1.5 and 3.0 percent, and the growth rate of the number of firms between around 1 percent.[24] The empirical results validate our quantitative finding of a positive local impact of migration; it further highlights the importance of incorporating the extensive margin of firm entry and exit into the analysis of migration, the central message of our paper.

Lastly, note that the per-capita GDP does not account for the cross-city differences in price levels in our data. However, this concern is partially alleviated since we use the first-difference between the year 2000 and 2005 as the dependent variable, and thus the time-invariant component of cross-city differences in price levels is already netted out. Nevertheless, the cross-city differences in price levels might change over time. In this regard, our model implies that the price level increases among cities that lose population but decreases among cities that gain population. The diverging change in the price level implies that the gaps in per-capita GDP growth rates between the migrant-sending and receiving cities should be even higher than what is observed in the data once the price levels are taken into consideration. This further implies that the positive relationship between the inflow rates and per-capita GDP growth rate in our paper is likely to be *underestimated* due to the omission of the price levels.

Aggregate Impacts of Migration Before we turn to the city-specific migration barriers, we briefly discuss the aggregate impacts of the existing migration flows between 2000 and

[24]Our point estimate on the elasticity between the inflow rate and per-capita GDP is smaller than the elasticity between the percentage of foreign-born population and per-capita GDP reported in Ortega and Peri (2014) (7.3 percent) in the context of cross-country migration using a gravity-based IV. In addition to the differences in estimation strategy, several other elements might also be responsible for the discrepancy between the estimates. First and foremost, the percentage of foreign-born population is a stock measure of migration which reflects the long-term impact, while the inflow rate is a flow measure over a short period of time. Moreover, cross-country migration might lead to higher returns from knowledge spillover due to the fact that the dispersion of human capital and technology is higher across countries than within a country.

2005. The aggregate results are reported in the last two rows in Table 6. At the aggregate level, the migration pattern described above increases the national real wage by 13.3 percent, which is about 25 percent of the economic growth in the data.[25] The gain in real wage comes from several channels. First, the marginal product of labor is usually higher in the larger cities, and thus, moving workers into those cities leads directly to real economic growth. Second, concentration of the population in the large cities increases consumption demand and lowers the marginal costs of production by reducing the costs of labor in the local markets. This leads to more firm entry in the 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 the other cities through intercity trade: the consumers in the other cities benefit from the decreased prices of the goods produced in the large cities, and the firms benefit from the lowered costs of intermediate goods. Similar to the local impacts, the surge in congestion dis-utility at the aggregate level also imposes a sizable cost, partially offsetting the aggregate gains in real wage. As a result, the national welfare only increases by 11 percent. The aggregate result in our paper is on par with Tombe and Zhu (2019) which finds a welfare gain of 8.5 percent. The similarity in aggregate welfare between our results and those in Tombe and Zhu (2019) suggests that the margin of firm entry is not quantitatively important at the aggregate level.

4.2 Migration Barriers and the Optimal City Size

After quantifying the local impacts of the migration flows, we turn to policy analysis and focus on city-specific migration barriers in the largest cities. We first discuss the magnitudes and implications of these migration barriers; we then study their local and aggregate impacts; lastly, we analyze the "optimal migration barriers" and the associated city size. The results presented in this section shed light on the welfare impacts of the policy-based internal migration restrictions, mainly the Hukou system. We focus on the largest four cities in China: Beijing, Shanghai, Guangzhou, and Shenzhen throughout this subsection, and show that relaxing the Hukou restrictions in these cities will not only bring about aggregate

[25]Over the same period, the real GDP of China increased by around 54 percent in the data according to Penn World Table 8.1.

economic gains but also increase the local welfare as well.

Our baseline estimation confirms that three out of the four cities, Beijing, Shanghai, and Guangzhou (hereafter "BSG") indeed enact additional migration barriers on top of the national average of $\bar{\lambda}$. The barrier to moving to Shanghai is 24 percent higher than the national average, followed by Guangzhou (13 percent) and Beijing (8 percent). These city-specific migration barriers are identified from the observed migration flows: the population inflows into the BSG are too small in the data, given the advantage in the real wage among BSG as compared to an average Chinese city in the baseline model. The differences between the model-implied and the observed population identify the magnitude of the additional entry barriers from the other parameters. The relative size of the migration barriers also speaks to the desirability of the cities as destinations: barriers into Shanghai (24 percent) being higher than into Beijing (8 percent) implies that Shanghai is a more attractive destination city. The relative advantage of Shanghai 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 broader base of potential migrants.

Removal of the additional migration barriers in Beijing, Shanghai, and Guangzhou always improves aggregate welfare, but not necessarily the local welfare. Setting the migration barriers in BSG to the national average level ($\delta_{(.)} = 1$) would lead to between 35.9 and 117.8 percent population increase in these cities.[26] The national welfare, as a result, increases by between 0.7 to 3.9 percent due to the additional benefits of population concentration. However, at the local level, the welfare impacts vary, as they depend on the changes in real wage v.s. congestion disutility. In the case of Beijing and Guangzhou, the local welfare increases after the removal of barriers by 9.7 and 18.6 percent, respectively. The residents in Shanghai, however, suffer a 8.3 percent welfare loss. The population of Shanghai would surge to 23 million if the additional barriers were removed. The resulting burden of congestion disutility exceeds the gain in the real wage, leading to a loss of the local welfare. The case of Shanghai reveals the potential conflict between the local and national welfare, which we

[26]Setting δ to 1 only decreases the migration barriers in these cities to the national average level. It does not imply frictionless migration.

will return to later in this section. The detailed results of these simulations are reported in Table 8.

Contrary to Beijing, Shanghai, and Guangzhou, the migration barriers in Shenzhen are roughly the same as the national average at 1.004. This is not a surprising result since Shenzhen is a city built from a tiny fishing village since the late 1970s, and it has been one of the few cities that actively encourages immigration into its jurisdiction. Setting δ to 1 in Shenzhen only slightly lower the migration barriers, and thus lead to a moderate response in the population (6 percent) and local welfare (1 percent) as reported in the last column of Table 8.

The largest cities are all underpopulated relative to the level of population that maximizes the national or the local welfare. To arrive at this conclusion, we simulate the model with different values of δ_i while keeping all the other parameters fixed. We compare the local and national welfare in city i across different values of δ_i to study the optimal δ_i and the associated population. Figure 8 reports the results by plotting the percentage change in welfare against the population in city i, which in turn depends on δ_i. The population that optimizes the welfare in Beijing is 18.1 million, which is 50 percent higher than the population in Beijing as of 2005 (12.6 million). If Beijing were to adopt this "local optima", the gain in the real wage would outweigh the loss in congestion disutility, leading to a welfare gain of 10 percent. However, the population of Beijing that maximizes the national welfare is significantly higher at 21 million. From the perspective of the central government, the productivity spillover from the "over-population" in Beijing exceeds the local costs of higher congestion, which justifies a higher population than the "local optima". The conflict of interest between the central and the local government highlights yet another controversial and under-explored dimension of migration issues. The results from the other cities are similar to those from Beijing, and we refer to readers to the others panels in Figure 8 for the details. The message of underpopulated Chinese cities confirms the findings in Au and Henderson (2006). Using a structural model, Au and Henderson (2006) show that between 50% to 60% of Chinese cities were underpopulated due to the restrictions in labor mobility, and the median loss of output is around 17% relative to the peak. Given the computational load, it is hard for our model to predict the fraction of underpopulated Chinese cities; however, we do find similar economic

implications of underpopulation in the largest cities.

4.3 Migration and Trade

In the last section, we study the interactions between migration and trade liberalization. We first show that the interaction between internal trade and migration is weak because the two serve as substitutes to each other. Internal trade liberalization benefits the smaller cities and reduce the need to migrate; at the same time, migration moves consumers closer to the production centers, and thus reduces the need for internal trade. We then move to international trade and argue that migration amplifies the gains from international trade by a wide margin.

Internal Trade Overall, the interaction between inter-city trade and migration is weak. We first show that inter-city trade only marginally affects migration. To do so, we perform a set of counter-factual analysis in which we first shut down the inter-city trade, and then lower $\bar{\lambda}$.[27] The results are reported in the third column of Table 6. Once the internal trade is shut down, the same reduction in $\bar{\lambda}$ leads to a 12.7 percent migration rate, which is 4.6 percentage points lower than the baseline estimation. The reduction in $\bar{\lambda}$, in turn, improves aggregate income by around 11.9 percent, which is only $1 - 11.9/13.3 \approx 11$ percent lower than the benchmark case with the intercity trade. The weak link between internal trade and migration is mainly due to the substitutability between the mobility of goods and people. Internal migration allows the consumers to live closer to the site of production, reducing the need for intercity trade. This substitutability implies that even if the intercity trade is completely absent, the gains from migration will only be marginally affected. Similarly, in the other direction, the volume of intercity trade is not responsive to migration frictions either. When we reduce the migration frictions in the baseline model, the overall trade openness *declined* slightly from 69 percent to 65 percent, again suggesting a weak interaction between internal trade and migration.

[27] We first set both $\bar{\lambda}$ and $\bar{\tau}$ to sufficiently high values such that no migration takes 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.

To further understand the effects gradual trade liberalization, we then 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.[28] 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 3.8 percent, as shown in Panel (a). All cities benefit from the internal trade liberalization mainly through two channels: 1) the direct benefit from the reduced transportation costs, which is the same as the "gains from trade" in a standard trade model, and 2) the indirect interactions between internal trade and migration. 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.[29] Internal migration plays a relatively minor role in amplifying the gains from internal trade liberalization: without any migration, the overall gain in the real wage is 3.4 percent, as shown in Panel (b) of the same figure. This implies that $3.4/3.8 \approx 89.5$ percent of the gain is through the first channel, while the other 10.5 percent is the amplification from the 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 only slightly from 82.7 to 82.3 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 Beijing is only 2.5 percent, while the smaller cities can enjoy gains at around 8.5 percent. This is an expected result from trade models following the work of Krugman (1980): small economies usually benefit more from trade liberalization because, after liberalization, the number of new imported varieties relative to the existing

[28] 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.

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

market size is more substantial 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 0.51 in the baseline model and 0.50 in the counter-factual. As a result, the need to migrate to large cities is mitigated in the first place.

The fact that the small cities benefit more than the large ones also implies that the 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 3.8 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 17.7 percent of the gain in real income was offset by higher congestion disutility, our results suggest that the reductions in internal trade frictions seem to be more "efficient" at the aggregate level.

International Trade Contrary to the previous exercise on the inter-city trade, migration greatly amplifies the gains from the international trade liberalization. To do so, we simulate a counter-factual world in which τ_{row} is lowered by 10 percent, while all of the other parameters are the same as in the baseline model. The results are reported in Figure 10.

Internal migration amplifies the gains from trade by around 55 percent. A 10-percent reduction in the international trade barrier leads to a 12.7-percent increase in the real income in the baseline model with migration. In contrast, without migration, the same reduction in trade barriers only leads to an 8.2-percent growth in real income, which implies that intercity migration amplifies the gains from trade by $12.7/8.2 - 1 \approx 55$ percent. The amplification works through the local wage rates. Reductions in trade barriers lead to the rapid expansion of the firms in coastal cities. However, higher labor demand quickly pushes up the wage rate in these cities, which in turn increases the marginal costs of production and throttles firm growth. Migration mitigates the surge of the local wage rates. With migration, higher wage rates in the coastal cities attract workers from the inland cities; the inflows of workers then shift the labor supply curve outward and dampen the equilibrium wage rate in the coastal cities. The additional labor supply brought by the migrants enables the exporting

firms to grow larger relative to the scenario without migration and eventually leads to a higher gains from trade. In the other direction, migration is also much more responsive to the international than the internal trade liberalization. The reduction in international trade barriers increases migration rate from 17.3 percent in the benchmark to 20.0 percent, leading to further gains in real wage and welfare due to the concentration of population as discussed in the previous section. The detailed results are reported in Figure 10.

Through the lens of international trade theory, almost all of the gains from trade come from the reallocation of resources within the country: the neo-classical trade models usually emphasize the inter-industry reallocation dictated by comparative advantage, and the new trade models following the work of Melitz (2003) highlight the gains due to the cross-firm reallocation. Our work 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 surge of the equilibrium wage rate, which usually works to limit the gains from trade, can be significantly mitigated. The resulting amplification effects can generate a gains from trade much higher than those measured in the standard trade models without migration. The spatial reallocation component also pushes our model outside of the definition of the Standard Trade Models as in Arkolakis et al. (2012), which implies that the amplification effects of migration cannot be captured merely 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 benefit the most from international trade liberalization. Panels (c) and (d) in Figure 10 highlight this pattern of welfare and real income gain. Despite the amplification effects of trade-induced migration, a higher concentration of population among the coastal cities takes a toll on the national welfare due to higher congestion disutility. National welfare only increases by 11.3, implying that around $1 - 11.3/12.7 \approx 11.0$ percent of the gain in the real wage is lost due to higher congestion disutility. The surge in the congestion disutility 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.[30] We decompose the gains from international trade into different channels in Table 9 to highlight the relative importance of the new elements in our model. Shutting down the migration channel reduces the gains from trade from 12.7 percent to 8.2 percent, and shutting down the firm entry reduces the gain to 6.0 percent. If we shut down both channels, the gains from trade in our framework are comparable to the majority of estimates in the literature, at around 5.5 percent. The fact that internal migration channel has such an enormous impact on the gains from trade is mainly due to its interaction with firm entry. Without the extensive margin, the gains from migration are significantly reduced compared to the benchmark exercise. Workers no longer anticipate more varieties and lower prices 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 18 percent population growth in Shenzhen; However, without firm entry, the same reduction in τ_{row} only increases the population in Shenzhen by 2.7 percent. The strong interaction between firm entry and migration also explains why shutting down both channels has similar effect as compared to only shut down the entry channel (5.5 percent v.s. 6.0 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 Extensions and Robustness Checks

5.1 Production Externality and City-Specific Productivity

Our baseline model abstracts away from the agglomeration forces within each city and the city-specific productivity. However, the concentration of workers and firms is known to gen-

[30]For example, the closest counterpart 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.

erate agglomeration externality through the Marshallian forces such as knowledge-spillover and labor market pooling (Rosenthal and Strange, 2004). Similarly, cities might also exhibit location-specific productivity due to the differences in resource endowment and existing infrastructure. In this section, we extend our baseline model to allow for such forces.

It is straightforward to modify the baseline model to capture these ideas. The starting point of the extension is a generalized production function similar to equation (4):

$$q_j^s(k) = \frac{A_j}{a(k)} b_j^s(k). \tag{10}$$

The new term, A_j, is the city-specific productivity that combines the location-specific productivity, $\bar{A}_j$, and productivity externality from agglomeration with an elasticity of γ:

$$A_j = \bar{A}_j \left(L_j\right)^{\gamma}.$$

The inclusion of the city-specific productivity modifies the input bundle requirement of all the firms in city j. Instead of using $a(k)$ units of input to produce one unit of output, now the firm needs to use:

$$\widetilde{a}_j(k) = \frac{a(k)}{A_j}.$$

In our baseline model, the distribution of $a(k)$ is the same across cities following $G(a)$. In the extension, one can directly show that the distribution function of $\widetilde{a}_j$ is now city-specific:

$$\begin{aligned} G_j(x) &= \Pr(\widetilde{a}_j \leq x) = \Pr(a \leq A_j x) = G(A_j x) = \left(A_j \mu\right)^{\theta} x^{\theta} \\ &= \mu_j^{\theta} x^{\theta}, x \in \left[0, \frac{1}{A_j \mu}\right]. \end{aligned}$$

The productivity distributions across cities are still Pareto with the same tail index, but instead of a common location parameter μ, we now have the city-specific location parameters, $\mu_j = A_j \mu$. If $A_i < A_j$, then the unit cost distribution in city i first-order dominates that in city j, capturing the idea that the firms in city i are more likely to receive a higher input requirement draw and thus a lower productivity. All the other parts of the model remain

unaffected up to the change from $a(k)$ to $\widetilde{a}_j(k)$.

To quantify the extension, we set $\gamma = 0.033$, the value from Au and Henderson (2006). To estimate $\bar{A}_j$, our starting point is the model-implied trade-balance condition in equilibrium:

$$(w_j)^{\varepsilon} = (L_j)^{\theta\gamma-1} \left(\bar{A}_j\right)^{\theta} \mathrm{MA}_j,$$

where MA_j summarizes the market access from city j to the rest of the economy in a similar manner to Donaldson and Hornbeck (2016):

$$\mathrm{MA}_j = \sum_{i=1}^{J} \frac{X_i^T}{\left(P_i^T\right)^{1-\varepsilon}} \left(\frac{\varepsilon}{\varepsilon-1}\tau_{ij}\right)^{1-\varepsilon} \frac{\left(a_{ij}^T\right)^{\theta-(\varepsilon-1)}}{\theta-(\varepsilon-1)}.$$

Taking logs on both sides of the equation, it is clear that $\bar{A}_j$ can be backed out as the residual of the following regression:

$$\log\left(w_j\right) = \beta_0 + \frac{\theta\gamma-1}{\varepsilon}\log\left(L_j\right) + \log\left(\mathrm{MA}_j\right) + \frac{\theta}{\varepsilon}\log\left(\bar{A}_j\right).$$

As a_{ij}^T is unobservable in the data, we instead approximate the market access as:

$$\widetilde{\mathrm{MA}}_j = \sum_{i=1}^{J} Y_i \left(\tau_{ij}\right)^{1-\varepsilon},$$

where Y_i is the output in city i and τ_{ij} is the iceberg trade costs that we have estimated earlier. The two other variables, w_j and L_j, are directly taken from the data. Denote the residual from the above regression as $\bar{u}_j$, we back-out $\bar{A}_j$ as:

$$\bar{A}_j = \exp\left(\frac{\varepsilon}{\theta}\bar{u}_j\right).$$

In the last step, we normalize the $\bar{A}_j$ vector so $\bar{A}_j$ in Shanghai is 1.

5.2 Results

We carry out two extensions along this direction. In the first extension, we set $A_j = (L_j)^{\gamma}$ so that the city-level productivity only depends on the agglomeration forces. In the second extension, we allow for both agglomeration and city-specific productivity, so that $A_j = \bar{A}_j (L_j)^{\gamma}$. In both extensions, we re-estimate all the other parameters following the same strategy as outlined in Section 3. Within each extension, we again compare the simulations with and without internal migration to study its impacts. The results are shown in the fourth and the fifth columns in Table 6 along side with the baseline results.

The main result of the baseline model is *strengthened* with the inclusion of the city-specific productivity and agglomeration: the local impacts of migration on real wage and welfare in the extensions are higher than in the baseline case. In both extensions, all the cities that receive net population inflow end up with higher real wage, the same as in the baseline model. Moreover, the average increase in the real wage in the extensions (11.9 and 17.3 percent) is significantly higher than in the baseline (4.7 percent). The higher local impacts are expected because 1) the concentration of population into the large cities now in itself leads to higher productivity, and 2) larger cities on average have higher $\bar{A}_j$ as well. The stronger impact in productivity is also reflected at the aggregate level as shown in the last row: the extended-model implies a significantly higher gain in the national-level income (30 and 38 percent) as compared to the baseline (13 percent). In the case where $A_j = \bar{A}_j (L_j)^{\gamma}$, 2 out of the 20 cities (Beijing and Shanghai) with a net inflow of population end up with lower local welfare due to the surge in congestion disutility. Beijing and Shanghai are significantly more productive than the national average as measured in $\bar{A}_j$, making them particularly attractive for migrants. As a result, the population in these two cities surge, leading to lower welfare in the end. Nevertheless, all the other receiving cities enjoy higher local welfare after taking congestion into account, and the average utility in all the 20 destination cities increases by 4.5 percent, again exceeding the welfare gain in the baseline case of 4.0 percent.

Higher Elasticity of Substitution Our last robustness exercise experiments with a higher elasticity of substitution, ε. As the "love-of-variety effect" drives the welfare impacts in our baseline model, one might expect that as the elasticity of substitution increases,

the extensive margin will not be as important. To explore this idea, we increase ε from the baseline level of 6 to a higher value of 8 and re-estimate the model following the same strategy as in the baseline model. The last column in Table 6 reports the results.

As expected, higher ε weakens the impacts of population flow both locally and nationally. While all the 48 cities with net inflow enjoy higher real wage and welfare, the magnitude (2.8 percent) is lower than in the baseline in both measures. At the national level, the increase in total income (6.6 percent) is also only half as compared to the baseline case. This exercise again highlights the importance of the extensive margin: if varieties are highly substitutable, the introduction of new firms and varieties following population inflow is no longer as desirable as before.

6 Conclusion

In this paper, we show that the local impacts of internal migration are positive in the destination cities in China. To do so, we extend a general equilibrium trade framework with endogenous firm entry and migration decision. The quantitative results suggest that all the cities that receive migrants benefit from the population inflow despite the negative impacts on congestion disutility. The key driving force is the margin of firm entry, without which inflows of workers would lower the real wage and welfare in the destination cities. Geography also plays an important role in shaping the impacts of internal migration: it is not the largest "super-star" cities that benefit the most from migrants (albeit they do benefit); surprisingly, it is the neighboring cities or the cities with the best market access to the national market that stand to win. These cities enjoy the productivity boom in the super-star cities through inter-city trade, and at the same time, they are less affected by the negative impacts resulted from the surging population than those "super-star" cities themselves. We uncover the geographical dimension of migration via estimating the model by taking into account the real-world transportation infrastructure, which also contributes to the under-explored literature in the context of China.

In addition to the main message, we also argue that the largest cities in China are underpopulated relative to the level that would maximize local or national welfare. On

average, the central government would prefer a higher population share in the largest cities. Finally, we show that the internal trade liberalization and migration serve as substitutes to each other, and migration amplifies the gains from international trade as it provides sufficient labor supply in the coastal cities.

Certain caveats exist when interpreting our results. We have ignored the issue of remittance; an explicitly modeled agricultural sector is absent in the current setup; and similarly, we do not delve into the housing market. To what extent the omission of these elements can affect our conclusion is a question that we are not able to answer at this stage. The interactions between the rural sector, the housing market, and the internal migration open up many daunting and yet fascinating questions, especially in the context of contemporary China. As tempting as it might be, we are not able to pack all these channels in our current setup, and we inevitably have to relay these crucial features, along with many others, to the future work.

References

Akay, Alpaslan, Corrado Giulietti, Juan D. Robalino, and Klaus F. Zimmermann, "Remittances and well-being among rural-to-urban migrants in China," *Review of Economics of the Household*, Sep 2014, *12* (3), 517–546.

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

_ , _ , and Yuta Takahashi, "Universal Gravity," Working Paper 20787, National Bureau of Economic Research December 2014.

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.

Artu, Erhan, Shubham Chaudhuri, and John McLaren, "Trade Shocks and Labor Adjustment: A Structural Empirical Approach," *The American Economic Review*, 2010, *100* (3), 1008–1045.

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

Brandt, Loren, Chang-Tai Hsieh, and Xiaodong Zhu, "Growth and structural transformation in China," *China's Great Economic Transformation*, 2008, pp. 683–728.

Caliendo, Lorenzo, Fernando Parro, Esteban Rossi-Hansberg, and Pierre-Daniel Sarte, "The Impact of Regional and Sectoral Productivity Changes on the U.S. Economy," *The Review of Economic Studies*, 2018, *85* (4), 2042–2096.

_ , Maximiliano Dvorkin, and Fernando Parro, "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock," *Econometrica*, forthcoming.

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.

Donaldson, Dave and Richard Hornbeck, " Railroads and American Economic Growth: A Market Access Approach *," *The Quarterly Journal of Economics*, 02 2016, *131* (2), 799–858.

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

Fajgelbaum, Pablo D, Eduardo Morales, Juan Carlos Surez Serrato, and Owen Zidar, "State Taxes and Spatial Misallocation," *The Review of Economic Studies*, 09 2018, *86* (1), 333–376.

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.

IMO, *World Migration 2005 Costs and Benefits of International Migration*, Vol. 3, Academic Foundation, 2006.

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

Liang, Zai, Jiejin Li, and Zhongdong Ma, "Migration and Remittances: Evidence from a poor province in China," *Asian Population Studies*, 2013, *9* (2), 124–141.

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.

Rosenthal, Stuart S. and William C. Strange, "Chapter 49 - Evidence on the Nature and Sources of Agglomeration Economies," in J. Vernon Henderson and Jacques-Franois Thisse, eds., *Cities and Geography*, Vol. 4 of *Handbook of Regional and Urban Economics*, Elsevier, 2004, pp. 2119 – 2171.

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

Taylor, J.Edward, Scott Rozelle, and Alan deBrauw, "Migration and Incomes in Source Communities: A New Economics of Migration Perspective from China," *Economic Development and Cultural Change*, 2003, *52* (1), 75–101.

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

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

A Tables and Figures

	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 (6). The data counterparts are computed from the Chinese City Statistical Yearbooks and the Custom Dataset.

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	5.38
ε	elasticity of substitution	6.00

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 *Annual Surveys of Manufacturing Firms*, and the value of ε comes from di Giovanni and Levchenko (2012).

Name	Data	Model	Diff.
Num. Firms	8.44	8.54	1.17%
Trade Share	0.62	0.65	3.27%
Tail Index	1.03	1.04	0.46%
Stay Rate	0.88	0.83	-6.06%
Std. Stay Rate	0.09	0.08	-14.34%
Corr(log(pop) inflow)	0.36	0.39	7.79%
Stay Rate Top 10	0.98	0.98	0.38%
Stay Rate Top 20	0.97	0.97	0.20%
Stay Rate Top 40	0.95	0.94	-0.35%
Stay Rate Other	0.87	0.81	-7.19%
Std(SR) Top 10	0.02	0.02	-18.46%
Std(SR) Top 20	0.02	0.02	17.39%
Std(SR) Top 40	0.04	0.04	-0.99%
Std(SR) Other	0.09	0.06	-31.62%
(Export+Import)/GDP	0.59	0.56	-5.03%
ROW/China Size	21.32	22.02	3.27%
Inflow Rate Beijing	0.17	0.17	-0.41%
Inflow Rate Shanghai	0.14	0.14	-1.18%
Inflow Rate Guangzhou	0.15	0.15	1.12%
Inflow Rate Shenzhen	0.54	0.59	10.05%

Table 3: 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.

Para.	Value	S.E
$\kappa \times 1000$	1.796	0.095
f_e	2.562	0.217
$\bar{\lambda} \times 1000$	12.131	0.641
$\bar{\tau}$	2.220	0.077
ϕ	3.522	0.034
ρ	195.938	5.999
τ_{row}	0.094	0.003
A	0.119	0.002
$\delta_{Beijing}$	1.081	0.008
$\delta_{Shanghai}$	1.243	0.016
$\delta_{Guangzhou}$	1.132	0.006
$\delta_{Shenzhen}$	1.003	0.012

Table 4: Parameters, Estimated

Note: This table reports the results of the estimation using SMM. The standard errors are computed with 100-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 $\delta_{\{\cdot\}}$ are the city-specific migration barriers.

	κ	f_e	$\bar{\lambda}$	$\bar{\tau}$	ϕ	ρ	τ_{row}	A	δ_{beijing}	δ_{shanghai}	$\delta_{\text{guangzhou}}$	δ_{shenzhen}
Num. Firms	1.813	-3.742	-4.624	-3.092	5.695	-0.456	-2.807	0.428	-0.305	-0.296	-0.219	-3.605
Trade Share	-0.629	0.827	1.642	-1.276	-2.041	0.164	-0.812	0.135	0.124	0.104	0.046	1.266
Tail Index	-0.015	0.039	0.066	0.058	-0.052	0.004	0.048	-0.027	0.007	0.003	0.002	0.029
Stay Rate	-1.313	0.779	2.289	0.791	-1.644	0.130	0.730	-0.111	0.098	0.076	0.074	1.005
Std. Stay Rate	4.947	-3.706	-9.655	-4.247	6.734	-0.534	-3.869	0.593	-0.341	-0.303	-0.309	-4.070
Corr(log(pop) inflow)	-0.524	3.335	4.513	3.178	-11.931	0.932	3.328	-0.499	-0.611	-0.543	-0.500	10.046
Stay Rate Top 10	-0.133	0.025	0.160	0.011	-0.118	0.010	0.014	-0.002	0.005	0.014	0.002	0.076
Stay Rate Top 20	-0.229	0.058	0.296	0.042	-0.209	0.017	0.045	-0.007	0.018	0.020	0.007	0.127
Stay Rate Top 40	-0.460	0.166	0.661	0.137	-0.461	0.037	0.141	-0.021	0.039	0.037	0.018	0.269
Stay Rate Other	-1.487	0.903	2.620	0.924	-1.884	0.149	0.849	-0.129	0.110	0.084	0.086	1.155
Std(SR) Top 10	7.435	-2.384	-10.137	0.163	8.886	-0.718	-0.037	-0.018	-0.536	-1.080	-0.452	-4.423
Std(SR) Top 20	7.934	-2.691	-11.196	-2.183	6.916	-0.554	-2.275	0.343	-0.862	-0.713	-0.296	-3.426
Std(SR) Top 40	7.380	-3.133	-11.273	-2.952	6.740	-0.538	-3.020	0.461	-0.890	-0.574	-0.257	-3.519
Std(SR) Other	4.312	-3.417	-8.670	-4.170	5.877	-0.468	-3.820	0.589	-0.279	-0.329	-0.259	-3.473
(Export+Import)/GDP	-0.562	0.739	1.468	-1.644	-1.813	0.145	-2.456	0.396	0.126	0.092	0.034	1.111
ROW/China Size	-0.745	0.930	1.889	2.838	-2.269	0.182	2.285	1.318	0.107	0.124	0.071	1.427
Inflow Rate Beijing	5.035	-3.954	-9.956	0.474	6.794	-0.545	0.861	-0.134	-12.247	0.062	0.040	0.553
Inflow Rate Shanghai	3.485	-4.025	-7.983	-4.154	12.230	-1.072	-3.976	0.634	0.044	-9.430	0.028	0.361
Inflow Rate Guangzhou	5.596	-3.817	-10.748	-6.674	1.567	-0.108	-6.492	1.036	0.048	0.047	-13.396	1.026
Inflow Rate Shenzhen	3.523	-6.159	-11.097	-5.786	17.369	-1.350	-5.702	0.822	0.052	0.043	0.055	-12.840

Table 5: Elasticity Matrix around Baseline

Note: The table reports the elasticity between the moments and the parameters in our benchmark estimation. For example, the first element, 1.813, means that if κ increases by 1 percent, then the "Number of Firms" moment increase by around 1.813 percent. We approximate the elasticities by 0.1 percent local variations around the benchmark estimation.

	Baseline	No Firm Entry	No Internal Trade	$A_j = (L_j)^\gamma$	$A_j = \bar{A}_j (L_j)^\gamma$	$\varepsilon = 8$
# inflow cities	40	55	38	22	20	48
# inflow cities with higher w/p	40	11	38	22	20	48
# inflow cities with higher u	40	9	38	22	18	48
$\Delta \log(L)$, inflow cities	0.093	0.048	0.058	0.157	0.237	0.079
$\Delta \log(w/p)$, inflow cities	0.047	-0.004	0.045	0.119	0.173	0.028
$\Delta \log(u)$, inflow cities	0.040	-0.006	0.040	0.083	0.045	0.028
$\text{corr}(\Delta \log(L), \Delta \log(w/p))$	0.977	-0.969	1.000	0.952	0.954	0.985
$\text{corr}(\Delta \log(L), \Delta \log(u))$	0.986	-0.958	0.995	0.979	0.760	0.985
$\Delta \log$ (aggregate income)	0.133	0.030	0.119	0.304	0.380	0.066
$\Delta \log$ (aggregate welfare)	0.110	0.023	0.101	0.204	0.144	0.066
aggregate migration rate	0.173	0.126	0.127	0.164	0.139	0.175

Table 6: The Impacts of Internal Migration

Note: The table presents the impacts of migration across various models. The first column is the baseline quantification; the second column is the model without firm entry. The third column is the result with internal autarky. The next two columns introduce city-level agglomeration and productivity; the last column re-calibrates the model using a higher ε. "Aggregate Migration Rate" is one minus the aggregate stay rate as defined in the main text.

	OLS				IV			
	(1) $\widehat{\text{Ln}\left(\frac{\text{GDP}}{\text{Pop.}}\right)}$	(2) $\widehat{\text{Ln}\left(\frac{\text{GDP}}{\text{Pop.}}\right)}$	(3) $\text{Ln}(\widehat{\#.\text{firms}})$	(4) $\text{Ln}(\widehat{\#.\text{firms}})$	(5) $\widehat{\text{Ln}\left(\frac{\text{GDP}}{\text{Pop.}}\right)}$	(6) $\widehat{\text{Ln}\left(\frac{\text{GDP}}{\text{Pop.}}\right)}$	(7) $\text{Ln}(\widehat{\#.\text{firms}})$	(8) $\text{Ln}(\widehat{\#.\text{firms}})$
Inflow Rate	0.379***	0.477***	0.226***	0.060	1.566***	3.072**	0.995***	1.138**
	(0.093)	(0.129)	(0.056)	(0.062)	(0.452)	(1.199)	(0.241)	(0.529)
Ln(Initial Pop.)		0.162***		-0.007		0.365***		0.079
		(0.054)		(0.025)		(0.115)		(0.052)
Ln(Initial GDP)		-0.091**		0.050***		-0.418***		-0.087
		(0.040)		(0.018)		(0.145)		(0.068)
Remoteness		0.249		-0.139		-0.375		-0.383*
		(0.195)		(0.139)		(0.399)		(0.229)
Constant	0.461***	0.623	0.314***	-0.182	0.587***	5.184***	0.397***	1.706*
	(0.021)	(0.456)	(0.012)	(0.265)	(0.057)	(1.869)	(0.030)	(0.935)
N	255	255	278	254	255	255	278	254
F statistic (IV)					43.626	11.280	45.139	11.273

Table 7: Inflow Rates, Number of Firms, and Per-Capita GDP

Note: This table reports the results of regressing the inflow rates on the log-difference of per capita GDP and the number of firms. Initial population and GDP refers to the year 2000. The log-difference in per capita GDP is between the year 2000 and 2005, and the number of firms, between 2004 and 2008 economic census. The instrument variable in columns 5 to 8 is the model-based inflow rates.

	Beijing	Shanghai	Guangzhou	Shenzhen
Baseline Barrier ($\delta_{(.)}$)	1.080	1.242	1.133	1.004
Baseline Population (Mil.)	12.607	16.804	9.529	15.607
Counterfactural Population (Mil.)	17.129	22.970	20.750	16.587
Change in Population	35.87%	36.69%	117.75%	6.28%
Change in Real Wage	22.43%	20.58%	54.36%	3.63%
Change in Congestion	194.37%	200.71%	1450.34%	23.93%
Change in Local Welfare	9.74%	-8.28%	18.57%	1.04%
Change in National Welfare	1.52%	0.70%	3.87%	0.35%

Table 8: Removing City-Specific Entry Barriers

Note: This table reports the results of counter-factual simulations in which we remove the additional migration barriers in each city by setting $\delta_{(.)}$ in that city to the national average level of 1. Note that the exercise only sets the migration barriers to the national average; it does not remove all the migration barriers. All of the changes are reported as percentage changes.

	$\Delta \log$(aggregate income)	$\Delta \log$(aggregate welfare)
Baseline	0.127	0.114
No Migration	0.082	0.085
No Firm Entry	0.060	0.060
No Migration and Firm Entry	0.055	0.057

Table 9: Aggregate Gains from 10% Reduction in International Trade Barrier

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}.

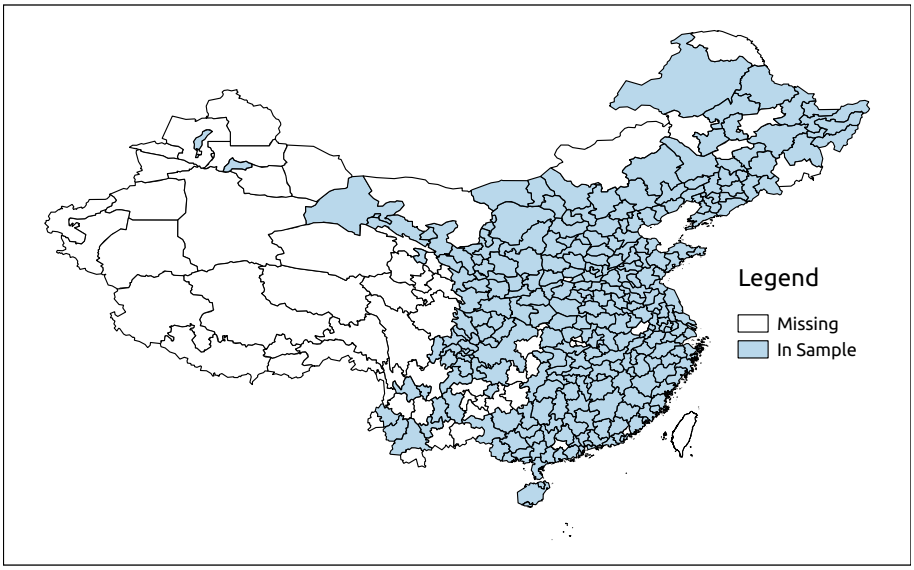
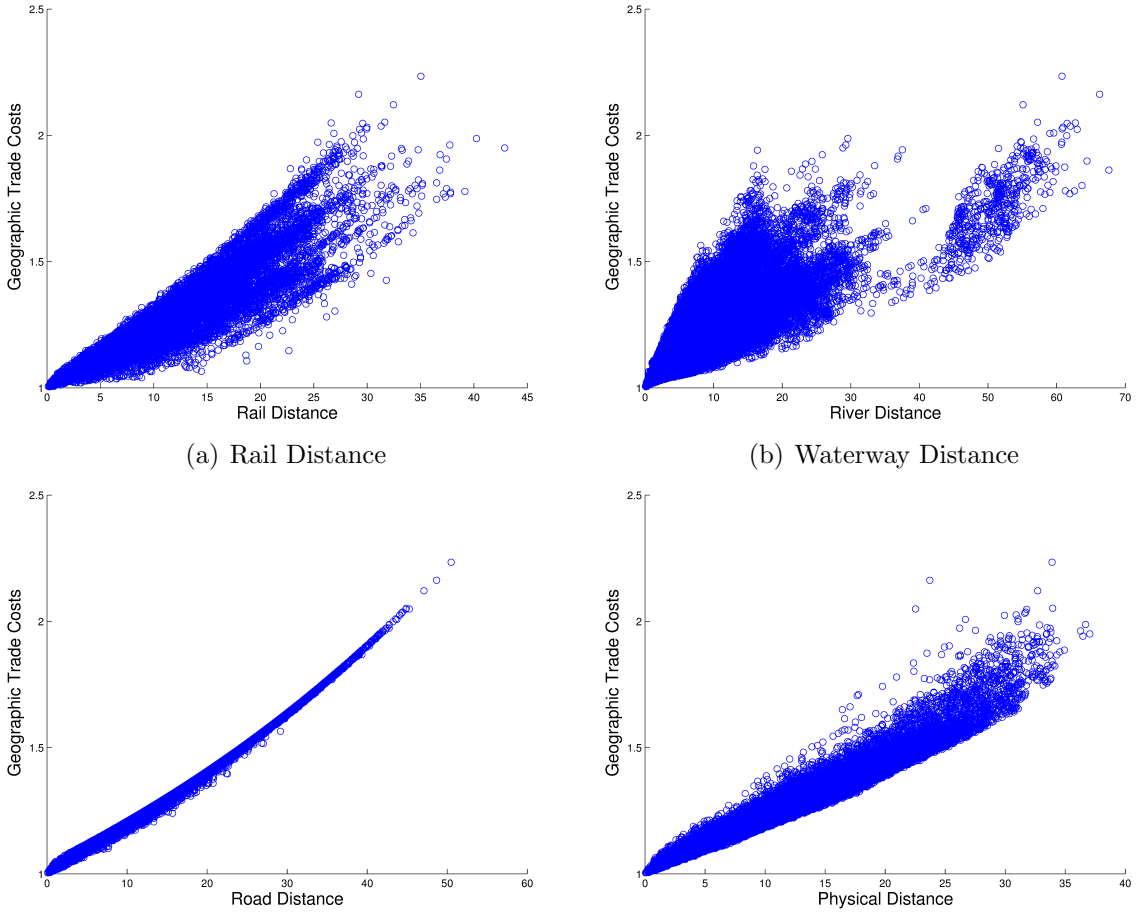


Figure 1: Prefecture-level Chinese Cities

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) Rail Distance

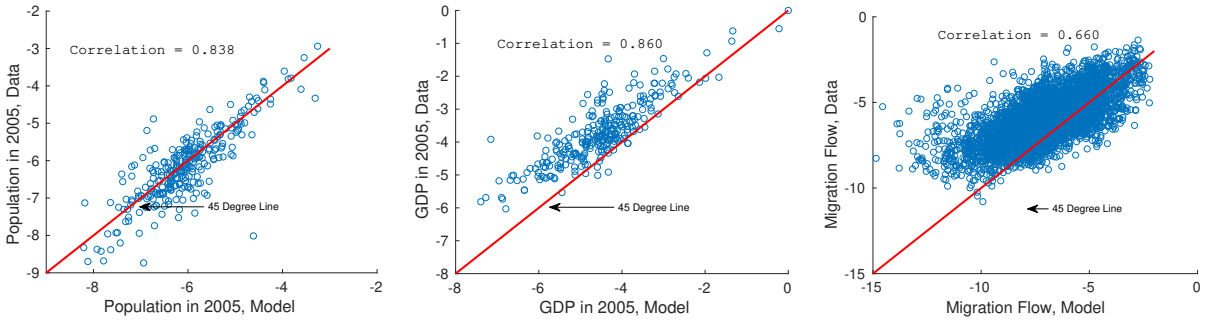
(b) Waterway Distance

(c) Road Distance

(d) Physical Distance

Figure 2: Geographic Trade Costs by Transportation Mode

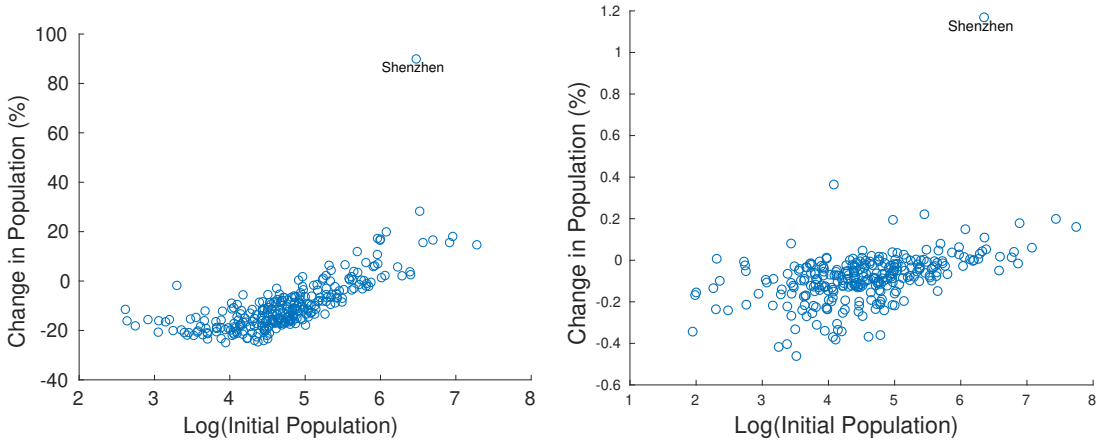
Note: The four panels above plot the estimated geographical costs matrix, T, against the mode-specific measures of distance obtained by the FMM. The last panel plots the T matrix against the physical distance between two cities. The physical distance is measured as the great circle distance between the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



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

Figure 3: Model Fit: Un-targeted Moments

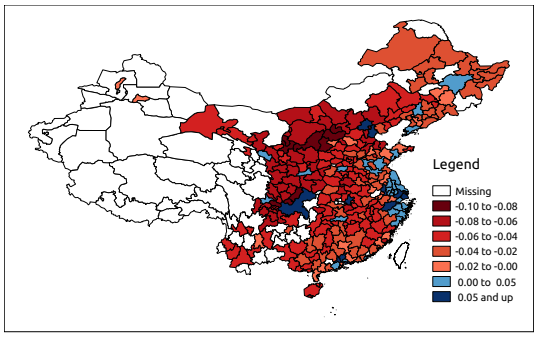
Note: The graphs above plot the population, GDP distribution, and bilateral migration flows implied by the model against their counter-parts in the data. In all of the graphs, the total population of China is normalized to be 1, and we plot the logarithms of the population and migration flows.



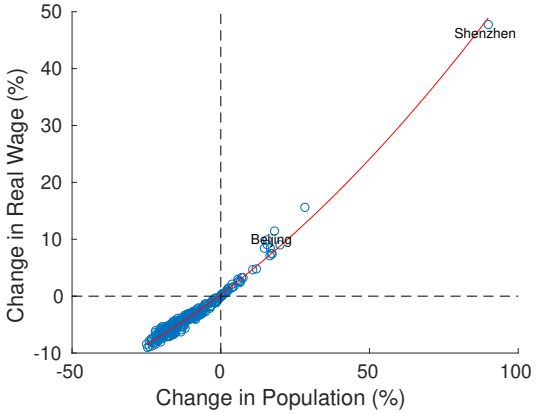
(a) Model (b) Data

Figure 4: Population Change and Initial Population

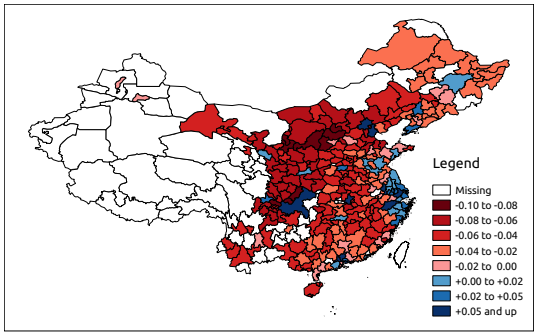
Note: The figure plots the change in population between 2000 and 2005 against the initial population in 2000. Panel (a) is our benchmark simulation and Panel (b) is the data.



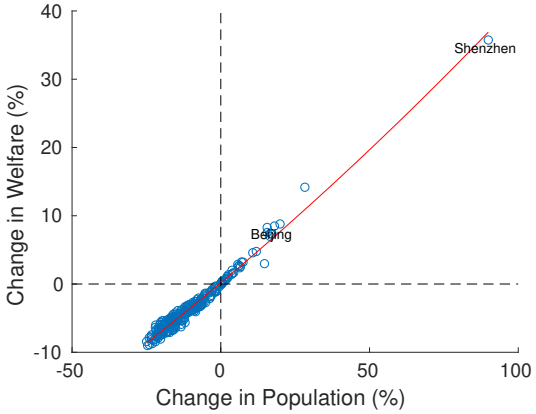
(a) Real Wage



(b) Population and Real Wage



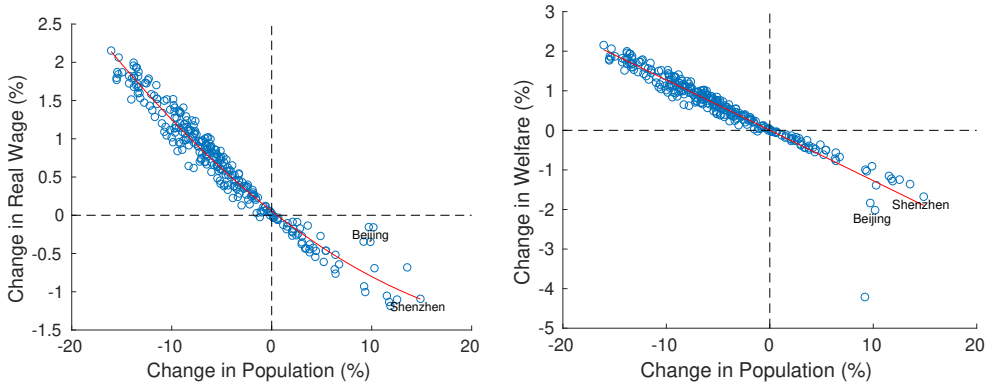
(c) Welfare



(d) Population and Welfare

Figure 5: Local Impacts of Migration

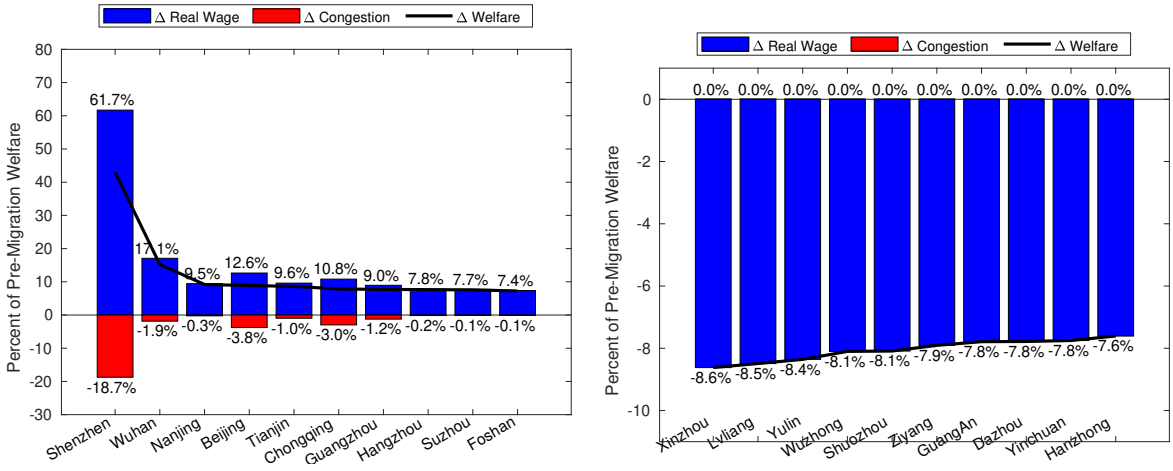
Note: The graphs above plot the change in real income and welfare between 2000 and 2005 due to inter-city migration implied by the model. The difference between real income and welfare is the congestion dis-utility. For more details, refer to Section 4.



(a) Real Wage (b) Welfare

Figure 6: Local Impacts without Firm Entry

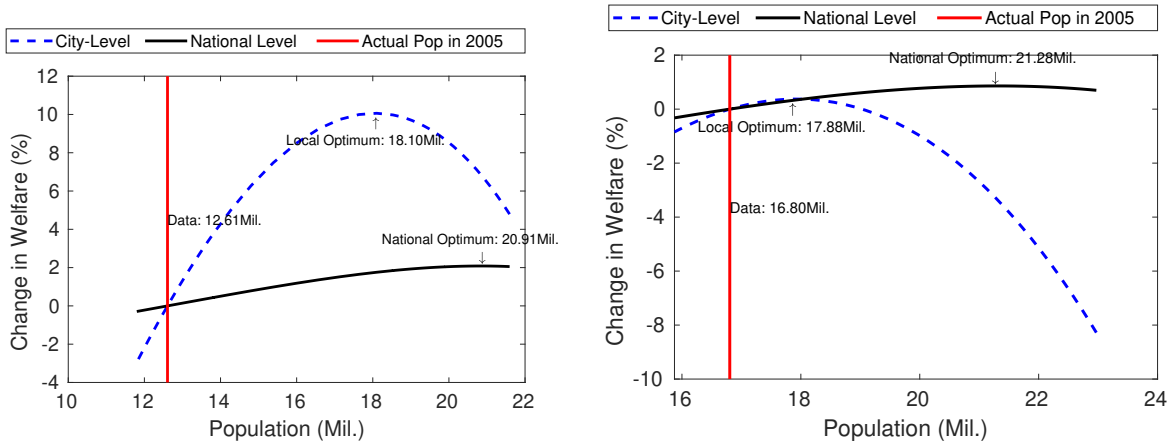
Notes: The graph plots the change in real income against the change in population in the counter-factual without firm entry. We fix the number of firms to the level in year 2000, and lower $\bar{\lambda}$ by the same magnitude as in Figure 5.



(a) Top 10 Winners (b) Top 10 Losers

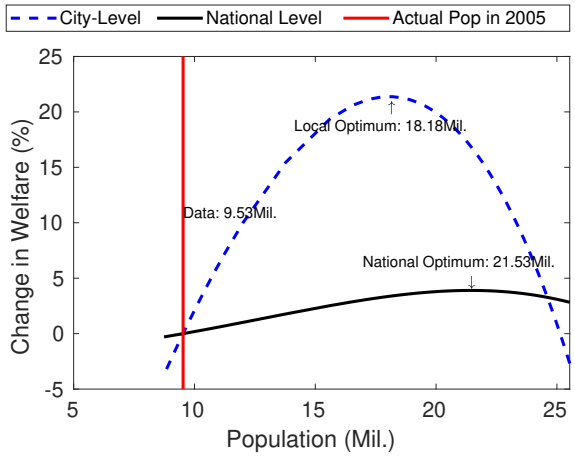
Figure 7: Winners and Losers from Migration

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.

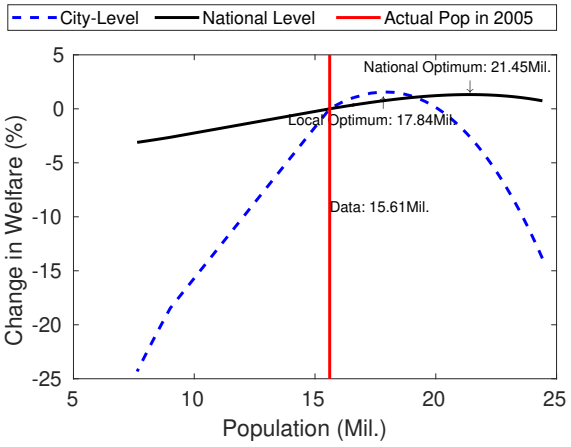


(a) Beijing

(b) Shanghai



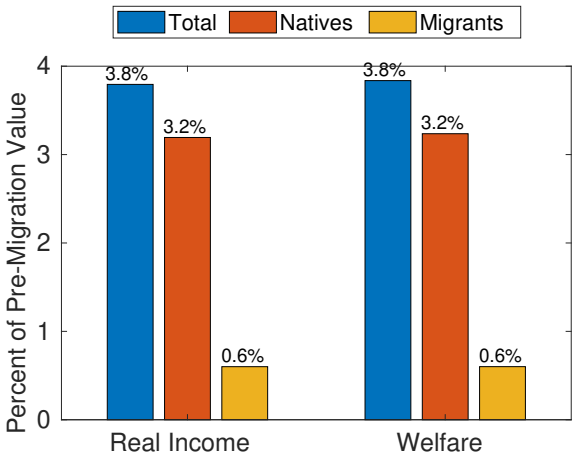
(c) Guangzhou



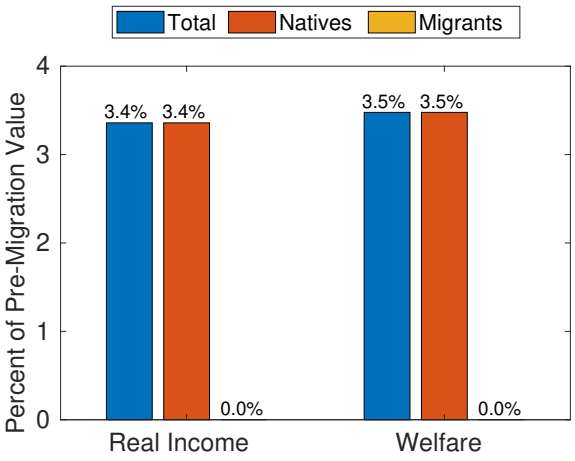
(d) Shenzhen

Figure 8: Optimal City Size at the Local and National Level

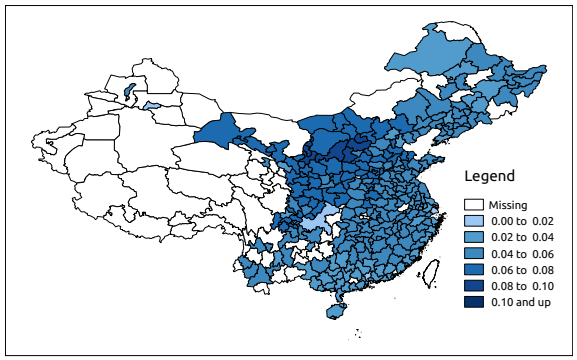
Note: This figure plots the local and national welfare of Beijing, Shanghai, Guangzhou, and Shenzhen as a function of population. The change in population is driven by the local migration barrier, δ_i. The welfare in the benchmark model (the red solid line) is normalized to 1.



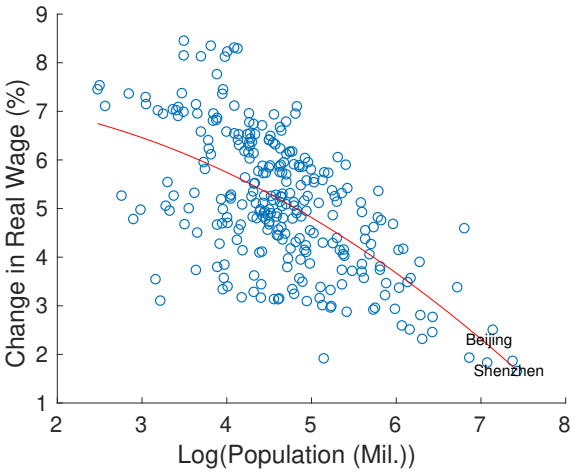
(a) Aggregate Impacts with Migration



(b) Aggregate Impacts without Migration



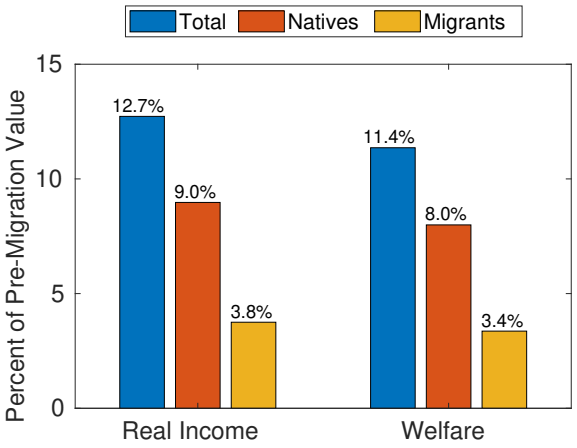
(c) Spatial Distributions of Welfare Gains



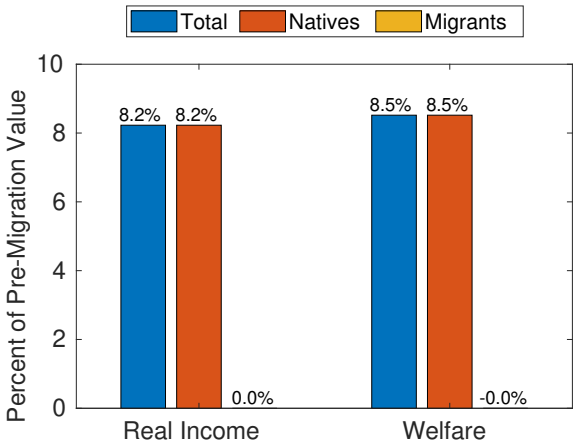
(d) Initial Population and Change in the Real Wage

Figure 9: Aggregate Impacts of Internal Trade Liberalization

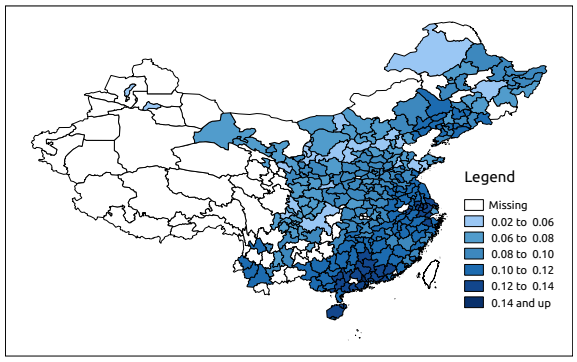
Note: The figures report the aggregate impacts and the direction of population flows from lowering the internal trade barrier, $\bar{\tau}$, by 10 percent while keeping $\bar{\tau} \cdot \tau_{\text{row}}$ the same as in the benchmark model.



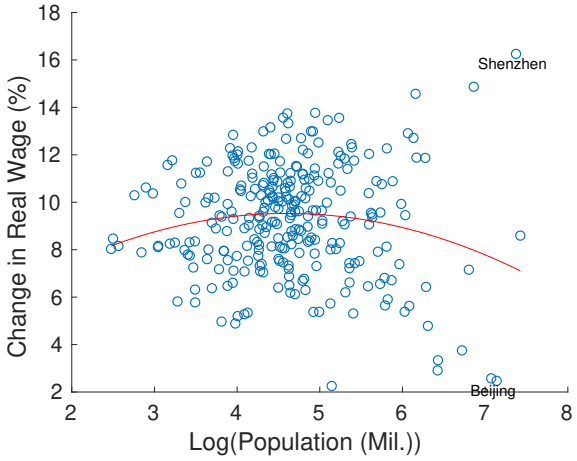
(a) Aggregate Impacts with Migration



(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.

B Solving the Model

In each city, we need to solve $\left\{w_j, I_j^N, I_j^T, P_j^T, P_j^N\right\}$. The free entry condition in each city and sector comes down to:

$$\frac{X_j^N}{\varepsilon\left(P_j^N\right)^{1-\varepsilon}}\left(\frac{\varepsilon}{\varepsilon-1}c_j^N\right)^{1-\varepsilon}\frac{\theta\mu_j^\theta\left(a_j^N\right)^{1+\theta-\varepsilon}}{\theta-(\varepsilon-1)} = c_j^N f_e + \mu_j^\theta\left(a_j^N\right)^\theta c_j^N f_{jj}$$
$$\sum_{i=1}^{J}\frac{X_i^T}{\varepsilon\left(P_i^T\right)^{1-\varepsilon}}\left(\frac{\varepsilon}{\varepsilon-1}\tau_{ij}c_j^T\right)^{1-\varepsilon}\frac{\theta\mu_j^\theta\left(a_{ij}^T\right)^{1+\theta-\varepsilon}}{\theta-(\varepsilon-1)} = c_j^T f_e + \sum_{i=1}^{C}\mu_j^\theta\left(a_{ij}^T\right)^\theta c_j^T f_{ij},$$

where a_j^N and a_{ij}^T are from zero-profit conditions:

$$a_j^N = \frac{\varepsilon-1}{\varepsilon}\frac{P_j^N}{c_j^N}\left(\frac{X_j^N}{\varepsilon c_j^N f_{jj}}\right)^{\frac{1}{\varepsilon-1}}$$
$$a_{ij}^T = \frac{\varepsilon-1}{\varepsilon}\frac{P_i^T}{\tau_{ij}c_j^T}\left(\frac{X_i^T}{\varepsilon c_j^T f_{ij}}\right)^{\frac{1}{\varepsilon-1}}.$$

The terms X_j^N and X_j^T are total expenditure, which are disciplined by the goods market clearing condition:

$$X_j^N = \alpha w_j L_j + \left(1-\beta^N\right)\eta^N X_j^N + \left(1-\beta^T\right)\eta^T X_j^T$$
$$X_j^T = (1-\alpha)\, w_j L_j + \left(1-\beta^N\right)\left(1-\eta^N\right) X_j^N + \left(1-\beta^T\right)\left(1-\eta^T\right) X_j^T.$$

Note that solving the system of equations together, it is straightforward to see that the sector-level expenditure are fixed fraction of total income, $w_j L_j$, in each city. The fixed fraction only depends on α, β^s, and η^s.

Under the assumption that $\frac{1}{a}$ follows a type-I Pareto distribution with the following CDF and PDF:

$$G_j(a) = \mu_j^\theta a^\theta$$
$$g_j(a) = \theta\mu_j^\theta a^{\theta-1}$$

we can explicitly derive the price index:

$$\begin{aligned}
P_j^N &= \left[\left(\frac{\varepsilon}{\varepsilon-1}c_j^N\right)^{1-\varepsilon} I_j^N \int_0^{a_i^N} a^{1-\varepsilon} g_j(a)da\right]^{\frac{1}{1-\varepsilon}} \\
&= \left[\left(\frac{\varepsilon}{\varepsilon-1}c_j^N\right)^{1-\varepsilon} I_j^N \theta\mu_j^\theta \int_0^{a_i^N} a^{\theta-\varepsilon} g(a)da\right]^{\frac{1}{1-\varepsilon}} \\
&= \left[\left(\frac{\varepsilon}{\varepsilon-1}c_j^N\right)^{1-\varepsilon} I_j^N \frac{\theta}{\theta-(\varepsilon-1)}\mu_j^\theta \left(a_j^N\right)^{\theta-(\varepsilon-1)}\right]^{\frac{1}{1-\varepsilon}} \\
&= \left[\left(\frac{\varepsilon}{\varepsilon-1}c_j^N\right)^{1-\varepsilon} I_j^N \frac{\theta}{\theta-(\varepsilon-1)}\mu_j^\theta \left(\frac{\varepsilon-1}{\varepsilon}\frac{P_j^N}{c_j^N}\left(\frac{X_j^N}{\varepsilon c_j^N f_{jj}}\right)^{\frac{1}{\varepsilon-1}}\right)^{\theta-(\varepsilon-1)}\right]^{\frac{1}{1-\varepsilon}}
\end{aligned}$$

$$\left(P_j^N\right)^{\frac{\theta}{\varepsilon-1}} = (\mu_j)^{\frac{\theta}{1-\varepsilon}} \left(\frac{\theta}{\theta-(\varepsilon-1)}\right)^{\frac{1}{1-\varepsilon}} \left(\frac{\varepsilon}{\varepsilon-1}\right)^{\frac{\theta}{\varepsilon-1}} \left(\frac{X_j^N}{\varepsilon}\right)^{\frac{\theta-(\varepsilon-1)}{(\varepsilon-1)(1-\varepsilon)}} \left[\left(c_j^N\right)^{-\theta} I_j^N \left(\frac{1}{c_j^N f_{jj}}\right)^{\frac{\theta-(\varepsilon-1)}{\varepsilon-1}}\right]^{\frac{1}{1-\varepsilon}}$$

and therefore:

$$P_j^N = \frac{\varepsilon}{\varepsilon-1}\left(\frac{\theta}{\theta-(\varepsilon-1)}\right)^{-\frac{1}{\theta}} \left(\frac{X_j^N}{\varepsilon}\right)^{-\frac{\theta-(\varepsilon-1)}{\theta(\varepsilon-1)}} \left[I_j^N \left(\frac{\mu_j}{c_j^N}\right)^\theta \left(\frac{1}{c_j^N f_{jj}}\right)^{\frac{\theta-(\varepsilon-1)}{\varepsilon-1}}\right]^{-\frac{1}{\theta}}$$

Similarly, the price index in the tradable sector is:

$$\begin{aligned}
P_j^T &= \left[\sum_{i=1}^J \left(\frac{\varepsilon}{\varepsilon-1}\tau_{ji}c_i^T\right)^{1-\varepsilon} I_i^T \int_0^{a_{ji}^T} a^{1-\varepsilon} g_i(a)da\right]^{\frac{1}{1-\varepsilon}} \\
&= \left[\sum_{i=1}^J \left(\frac{\varepsilon}{\varepsilon-1}\tau_{ji}c_i^T\right)^{1-\varepsilon} I_i^T \frac{\theta}{\theta-(\varepsilon-1)}\left(\mu_i\right)^\theta \left(a_{ji}^T\right)^{\theta-(\varepsilon-1)}\right]^{\frac{1}{1-\varepsilon}} \\
&= \left[\sum_{i=1}^J \left(\frac{\varepsilon}{\varepsilon-1}\tau_{ji}c_i^T\right)^{1-\varepsilon} I_i^T \frac{\theta}{\theta-(\varepsilon-1)}\left(\mu_i\right)^\theta \left(\frac{\varepsilon-1}{\varepsilon}\frac{P_j^T}{\tau_{ji}c_i^T}\left(\frac{X_j^T}{\varepsilon c_i^T f_{ji}}\right)^{\frac{1}{\varepsilon-1}}\right)^{\theta-(\varepsilon-1)}\right]^{\frac{1}{1-\varepsilon}}
\end{aligned}$$

$$\left(P_j^T\right)^{\frac{\theta}{\varepsilon-1}} = \left(\frac{\varepsilon}{\varepsilon-1}\right)^{\frac{\theta}{\varepsilon-1}} \left(\frac{\theta}{\theta-(\varepsilon-1)}\right)^{\frac{1}{1-\varepsilon}} \left(\frac{X_j^T}{\varepsilon}\right)^{\frac{\theta-(\varepsilon-1)}{(\varepsilon-1)(1-\varepsilon)}} \left[\sum_{i=1}^J I_i^T \left(\frac{\mu_i}{\tau_{ji}c_i^T}\right)^\theta \left(\frac{1}{c_i^T f_{ji}}\right)^{\frac{\theta-(\varepsilon-1)}{\varepsilon-1}}\right]^{\frac{1}{1-\varepsilon}}$$

and therefore:

$$P_j^T = \frac{\varepsilon}{\varepsilon - 1}\left(\frac{\theta}{\theta - (\varepsilon - 1)}\right)^{-\frac{1}{\theta}}\left(\frac{X_j^T}{\varepsilon}\right)^{-\frac{\theta-(\varepsilon-1)}{\theta(\varepsilon-1)}}\left[\sum_{i=1}^{J} I_i^T\left(\frac{\mu_i}{\tau_{ji}c_i^T}\right)^{\theta}\left(\frac{1}{c_i^T f_{ji}}\right)^{\frac{\theta-(\varepsilon-1)}{\varepsilon-1}}\right]^{-\frac{1}{\theta}}$$

Finally, trade balance requires $X_j^T = \sum_{i=1}^{J} X_{ij}^T$, so in each city j we have $w_j L_j$ equal to the following:

$$w_j L_j = \sum_{i=1}^{J} \frac{I_j^T \tau_{ij}^{-\theta} \left(f_{ij}\right)^{-\frac{\theta-(\varepsilon-1)}{\varepsilon-1}} \left(w_j^{\beta_T}\left[\left(P_j^N\right)^{\eta^T}\left(P_j^T\right)^{1-\eta^T}\right]^{1-\beta^T}\right)^{-\frac{\theta-(\varepsilon-1)}{\varepsilon-1}}}{\sum_{k=1}^{J} I_k^T \tau_{ik}^{-\theta}\left(f_{ik}\right)^{-\frac{\theta-(\varepsilon-1)}{\varepsilon-1}}\left(w_k^{\beta^T}\left[\left(P_k^N\right)^{\eta^T}\left(P_k^T\right)^{1-\eta^T}\right]^{1-\beta^T}\right)^{-\frac{\theta-(\varepsilon-1)}{\varepsilon-1}}} w_i L_i. \quad (11)$$

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 the wage rate in Shanghai 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 (8). 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 $W_{(.)}$ 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 can no longer be improved (stall counter exceeds stall limit).

C.2 Bootstrapping

We estimate the standard errors using a 100-repetition bootstrapping. In each repetition, we bootstrap the following data samples to re-compute the target moment in equation 8:

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 100 samples to generate the bootstrapped distribution of each estimated parameter. The standard errors of each parameter can be directly derived from the bootstrapped distribution.