Migration and Resource Misallocation in China

Xiaolu Li Lin Ma Yang Tang [*]

Nov 2023

Abstract

We structurally estimate the firm-level frictions across prefectures in China and quantify their aggregate and distributional implications. Based on a general equilibrium model with input and output distortions and migration, we show that the firm-level frictions are less dispersed and less correlated with firm productivity in richer prefectures. Counterfactual exercises show that reducing the within-prefecture misallocation increases aggregate welfare, discourages migration toward large prefectures, and reduces spatial inequality. Moreover, internal migration alleviates micro-frictions' impacts on aggregate welfare and worsens their effects on spatial inequality.

Keywords: misallocation; regional trade; economic geography; welfare gain

JEL Classification: F12;O11;R12

Published in Journal of Development Economics, 2024, 167, 103218.

DOI: 10.1016/j.jdeveco.2023.103218

[*] The paper formerly circulated under the title "Migration and Spatial Misallocation in China". Respectively: Nanjing University of Posts and Telecommunications (xiaoluleez@gmail.com); Singapore Management University (linma@smu.edu.sg); Nanyang Technological University (tangyang@ntu.edu.sg). The authors are especially grateful for valuable insights from the editor, two anonymous referees, and the stimulating discussions with Wen-Tai Hsu, Zheng Michael Song, Cheng Wang, and Jing Wu. We have also greatly benefited from the comments from various seminars and conferences at the Chinese University of Hong Kong, Nanyang Technological University, the Taipei International Conference on Growth, Trade, and Dynamics, the Econometric Society China Meeting, and ABFER Annual Meetings. We are solely responsible for any remaining errors.

1 Introduction

Frictions at the firm level are costly: they affect production decisions, distort resource allocation, and lower aggregate productivity. The existing studies on micro-level frictions mainly focus on cross-country differences and show that these frictions are rooted in institutional quality, geography, and infrastructure investment.[1] Similar to the international context, these underlying factors also vary substantially across regions within the same country. However, different from the international context, due to labor mobility within a country, the micro-frictions exert additional influence by reshaping the distribution of economic activities across space. The interaction between region-specific frictions and labor mobility implies that policies addressing misallocation in one location may generate positive or negative spillover effects on other regions through migration flows. To quantify the impacts of such frictions, one needs a general equilibrium framework with both micro-level frictions and labor mobility, and this paper aims to do so.

We accomplish three goals in this paper. First, we propose a general equilibrium framework that incorporates micro-level frictions following Hsieh and Klenow (2009) into a multi-region trade model that allows for migration, heterogeneous firms, and endogenous firm entry in the spirit of Melitz (2003). Second, we design an empirical strategy to structurally estimate the joint distribution of micro-frictions and firm-level productivity within each location. Lastly, we evaluate how micro-level frictions affect aggregate welfare, spatial distribution of resources, and their interaction with migration through counterfactual simulations in the context of China.

We classify the firm-level frictions into output and labor frictions. The output frictions are revenue wedges that distort the firm's profit margin, and the labor frictions are payroll wedges that affect the costs of hiring workers. In the model, each firm draws the two wedges and its productivity from a location-specific joint distribution and makes its sales decisions based on its draw. At the aggregate level, firm-level frictions manifest themselves through multiple channels. Within each location, the frictions affect the wage rates and price levels as they distort the firms' decisions in the factor and the output markets. Across locations, the

[1] For literature on micro-level frictions and resource misallocation, see Restuccia and Rogerson (2008), Hsieh and Klenow (2009), Buera et al. (2011) and Midrigan and Xu (2014), among others for examples.

micro-frictions permeate through inter-region trade as they distort the costs of intermediate inputs. Lastly, these wedges also influence migration decisions at the individual level through their impacts on wage rates and price levels.

Based on this model, we design an empirical strategy to structurally estimate the joint distribution of the firm-level frictions and productivity in each location. Estimating the joint distribution in the literature is challenging because we only observe a subset of these firms in the data due to selection.[2] Failing to account for selection leads to biased results, as noted in Bai et al. (2019) and Yang (2021). Structurally estimating the distribution by simulating the general equilibrium addresses the selection issue; however, general equilibrium is computationally costly to evaluate even with a single location, as seen in Bai et al. (2019). Structural estimation by solving the general equilibrium is prohibitively expensive with multiple locations and factor mobility. We thus rely on the Simulated Method of Moments (SMM) to overcome this difficulty. In particular, we use the sampling scheme in the SMM to approximate sample selection. The approximation allows us to estimate the joint distribution for many locations without solving for the general equilibrium.

Based on the procedure described above, we estimate the joint distributions of productivity and frictions in 237 prefectures in China from 1998 to 2007. The firm-level data come from the *Annual Surveys of Industrial Firms* ("*ASIF*" thereafter) conducted by the National Bureau of Statistics (NBS). We use each firm's observed sales, payroll, and total production costs to estimate the firm-level frictions following Hsieh and Klenow (2009) and then use the SMM to infer the parameters governing the joint distributions. The parameters of interest are the standard deviations of the frictions and the pair-wise correlations among firm-level productivity, output friction, and labor friction within a location.[3]

Two distinct patterns emerge from the estimation results. First, while both output and labor frictions are prevalent in all the prefectures, the dispersion of labor frictions is substantially higher. The average standard deviation of output frictions is 0.11, and the same statistics for labor frictions are one order of magnitude higher at 1.18. The large

[2] Selection in the data could come from either survival attrition or size thresholds on sample selection.

[3] Our baseline estimation assumes that within each location, the unconditional mean of the marginal distribution of firm-level frictions is zero. Therefore, frictions in the main text result in resource misallocations within-location, not between-location. Appendix D presents results with both within- and between-location resource misallocation.

standard deviation in the labor frictions indicates that the input markets are more distorted than the output markets. Second, the frictions are less dispersed and less correlated with productivity in prefectures with higher per-capita GDP. For example, while the average dispersion of the output frictions is around 0.085 among the coastal prefectures, the same statistics is around 0.108 for the inland and poorer regions. These findings suggest that part of the observed income disparity and population distribution might be attributed to the spatial distribution of firm-level frictions: our reduced-form analyses show that around 13% of the observed variations in per-capita GDP across prefectures can be explained by the dispersion and productivity correlations of the frictions.

We perform counterfactual simulations to evaluate the aggregate and distributional implications of the observed micro-friction distributions. In the first exercise, we reduce the standard deviations of the output and labor frictions in all the prefectures by 0.01. Reducing the dispersion of frictions leads to a 3.19% increase in aggregate welfare, corresponding to a semi-elasticity of -3.19. When the distributions of frictions become less dispersed, the firms are more likely to draw frictions closer to zero, which improves aggregate welfare through several channels. As the labor frictions converge towards zero within each prefecture, the marginal product of labor is more equalized across firms; similarly, the reduction in the dispersion of output frictions better aligns market share with firm-level productivity. Both forces lead to aggregate gains in output. Reducing the dispersion of firm-level frictions also decreases spatial inequality. With lower standard deviations of friction, individuals are less likely to migrate to rich prefectures and the Gini coefficient of real wage declines. The equalizing effect comes from the fact that smaller and poorer prefectures facing higher dispersion in the data. When the distribution becomes less dispersed, the smaller and poorer prefectures benefit more, encouraging more firm entries and attracting higher population inflow.

Migration interacts with the micro-frictions in a meaningful way: migration *alleviates* the impacts of micro-frictions on the aggregate welfare and *amplifies* its distributional consequences. To show this, we shut down migration in the baseline model and repeat the exercise of reducing the dispersion parameters of frictions. Without internal migration, the aggregate welfare gain from the same 0.01 reduction in the standard deviation of both frictions is 3.37%, higher than the impacts with internal migration. In the model with migrations, the negative impacts of frictions at the national level are partially offset by the migration flows — people leaving the heavily distorted regions in favor of the less distorted ones, therefore reducing the economic activity at the heavily distorted locations. Without migration, the aggregate impacts of friction are higher as people can no longer escape the heavily distorted prefectures.

The above mechanism also implies that spatial inequality becomes less responsive to micro-frictions when migration is shut down. For example, lowering the dispersion parameters reduces the Gini coefficient by 0.3 with migration, but the impact reduces to 0.25 without migration. Intuitively, while a less dispersed distribution of frictions lowers spatial equality, migration works to amplify this effect by moving people out of those heavily distorted regions, which tends to improve the real wage at the left tail.[4]

Lastly, we show that while both "migration liberalization" and "reduction in within-prefecture misallocation" improve aggregate welfare, their implications on spatial inequality are drastically different. To highlight this point, we simulate another counterfactual in which the overall migration barriers decline to achieve the same 3.19 percent welfare gain as the baseline exercise. While both policies lead to the same welfare gain, migration liberalization worsens spatial inequality because the migrants prefer to move towards the richer and less-distorted prefectures once the migration barriers decline. The population concentration in these large prefectures further increases their productivity advantages through agglomeration, leading to higher spatial inequality. On the contrary, when the friction distribution becomes less dispersed, the productivity gaps between the poor and rich regions shrink, resulting in lower spatial disparities.

Our paper is closely related to the literature on micro-level frictions and resource misallocation (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009; Buera et al., 2011; Guner et al., 2008; Hopenhayn, 2014; Hsieh and Moretti, 2019). In the context of China, Brandt et al. (2013) measure the reduction in the aggregate non-agricultural TFP due to labor and capital distortions across provinces and sectors for the period 1985-2007. Brandt et

[4] As it will be clear later in this paper, we use a Melitz (2003) framework that features many agglomeration forces in the New Economic Geography literature (Krugman, 1991; Fujita et al., 1999). These forces originate from the increasing returns to scale production function, monopolist competition, and trade costs.

al. (2017) emphasizes the role of entry barriers in regional income growth. Song and Wu (2015) and Wu (2018) focus on the capital misallocation behind the micro-frictions among Chinese firms. We contribute to this literature in two ways: we are the first to incorporate endogenous migration into a general equilibrium framework with the micro-frictions. We are also the first to propose a structural estimation strategy that can be efficiently implemented in settings with many locations.

Our work is also broadly related to the literature on the Chinese economy. Brandt et al. (2008) document the process of industrial transformation, the role played by institutions and barriers to factor allocation. Song et al. (2011) argue that the reduction in the distortions associated with state-owned enterprises may be responsible for the rapid economic growth since 1992. Hsieh and Song (2015) use firm-level data to show that the reforms of the state sector were responsible for 20 percent of aggregate TFP growth from 1998 to 2007. Tombe and Zhu (2019) study how goods and labor market frictions affect aggregate productivity at the province level. We show that the spatial dispersion of frictions lowers aggregate welfare and output, and contributes to regional income differences. Moreover, from the policy perspective, we highlight that while both migration liberalization and reduction in frictions improve aggregate welfare, the former increases, but the latter decreases spatial inequality.

The rest of this paper is organized as follows: Section 2 presents the theoretical framework and defines the frictions. We discuss the estimation strategies and results in Section 3. Section 4 presents the calibration strategy, and Section 5 describes the quantitative results. Section 6 concludes.

2 The Model

We introduce labor and output frictions following Hsieh and Klenow (2009) into a multi-location framework with migration, internal trade, heterogeneous firms, and endogenous firm entry decisions similar to Ma and Tang (2020).

2.1 Basic Setup

Consumers The economy contains J geographically segmented locations in the set $\mathcal{J} = \{j|j = 1, 2, \cdots, J\}$. We index the locations with either i or j. Workers are identical and mobile across locations subject to migration frictions. Workers residing in location j obtain utilities from consuming the set of varieties available in location j:

$$U_j = \left[\sum_{k \in \Theta_j} y\left(k\right)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon}{\varepsilon-1}} \left(L_j\right)^{\psi} \phi_j, \tag{1}$$

where ε represents the elasticity of substitution across all varieties, $y(k)$ denotes the consumption of variety k, and Θ_j denotes the set of available varieties in location j. The last term, $\left(L_j\right)^{\psi} \phi_j$, measures the amenity in location j, and ψ captures the extent to which amenities are affected by the population, and ϕ_j is the fundamental amenity of the location.

Firms The production side follows Melitz (2003) with heterogeneous firm-level productivity and endogenous firm entry.[5] The heterogeneity in productivity allows us to study the correlation between frictions and productivity, and endogenous firm entry allows the frictions to affect aggregate welfare through the number of varieties available to consumers. Each variety k is produced by a unique firm using local labor and input bundles. The market structure is monopolistic competition. Firms are heterogeneous in their productivity. The production function is:

$$q_j(k) = A_j \times a(k) \times b_j(k), \tag{2}$$

where $q_j(k)$ is the output of variety k, $a(k) > \underline{a}$ is the firm-level productivity, and A_j is the location-specific productivity. The last term in equation (2), $b_j(k)$, is the input bundle for

[5] We follow Melitz (2003) to introduce two forces in the model: the interaction between productivity and micro-level frictions, as well as endogenous firm entry that is quantitatively important as shown in Section 5 later.

the production of variety k, which can be produced according to:

$$b_j(k) = \left[\ell_j^b(k)\right]^{\beta_j} \left[\sum_{k' \in \Theta_j} \left(y^b\left(k'; k\right)\right)^{\frac{\varepsilon-1}{\varepsilon}}\right]^{\frac{\varepsilon(1-\beta_j)}{\varepsilon-1}}.$$

The variable $\ell_j^b(k)$ denotes the labor input, and $y^b(k'; k)$ is the demand for variety k' to produce input bundle for the production of the variety k. The parameter $0 < \beta_j < 1$ denotes the labor share in the production.[6]

Each variety is tradable across locations, subject to costs. Standard iceberg trade costs also apply: to deliver 1 unit of variety from j to i, firms must produce and ship $t_{ij} \geq 1$ units from j. We assume $t_{jj} = 1, \forall j \in \mathcal{J}$. In the baseline model, we do not assume any fixed costs of production or exporting for simplicity. We explore the effects of these fixed costs and show in Section 5 that our results are robust to including fixed costs.

In addition to the heterogeneity in productivity, firms are subjected to two types of idiosyncratic frictions: output and labor frictions. The output frictions, $\tau_y(k) < 1$, are equivalent to the revenue wedges such that the firm only receives a fraction $[1 - \tau_y(k)]$ of its revenue, $r_j(k)$. The output frictions can be negative to reflect a production subsidy. The labor frictions, $\tau_\ell(k) > -1$, are the frictions that affect the marginal product of labor relative to the composite varieties. A firm needs to pay $[1 + \tau_\ell(k)]$ times the local wage rate, w_j, to hire one unit of labor. The labor frictions can also be negative, implying payroll subsidies. The three firm-level shocks, $\{a(k), \tau_y(k), \tau_\ell(k)\}$, are drawn from a location-specific joint distribution with the cumulative distribution function $G_j(a, \tau_y, \tau_\ell)$.

We interpret both frictions as payroll or output "taxes" and "subsidies" implemented by the local government and use these terms interchangeably with "frictions" in the rest of the paper. We also assume that the local governments maintain balanced budgets. Thus, any surplus (deficits) from the distortionary taxes (subsidies) will be offset by rebates to or lump-sum taxes on the local population. We denote the per-capita transfer in location j as γ_j.

Lastly, we allow for endogenous entry of the firms. Infinitely many potential entrants

[6] Different from the standard neoclassical production function, we omit the capital inputs and instead assume capital goods to be a subset of available varieties in the economy.

with a zero outside option reside in each location. A potential entrant in location j can pay a cost of f_e in the unit of the local input bundle to start producing a new variety.[7] Denote the number of entrants in location j as I_j. Upon entry, the firm draws its productivity and frictions from the joint distribution $G_j(a, \tau_y, \tau_\ell)$ and starts production.

Migration Workers can migrate between the locations subject to costs. The migration decision hinges on three elements: the indirect utility, the idiosyncratic preference, and the bilateral migration costs. The indirect utility of living in location j, which we denote as V_j, can be expressed as:

$$V_j = \frac{w_j + \gamma_j}{P_j} \left(L_j\right)^{\psi} \phi_j, \tag{3}$$

where w_j is the wage rate, γ_j is the transfer payment (tax), P_j is the ideal price index, and $\left(L_j\right)^{\psi} \phi_j$ denotes the (endogenous) amenity in location j. In addition, each worker draws an idiosyncratic preference shock toward each location $\{\nu_j\}_{j=1}^{J}$, where ν_j is *i.i.d* across locations and individuals. We assume that ν_j follows a Frechet distribution with the CDF:

$$F(\nu) = \exp\left(-\nu^{-\kappa}\right),$$

where κ is the shape parameter, which measures the elasticity of bilateral migration with respect to frictions. Lastly, moving from location j to i also incurs an origin-destination-specific dis-utility, which are denoted as $\lambda_{ij} \geq 1$, with $\lambda_{jj} = 1, \forall j \in \mathcal{J}$. The costs of migration capture not only the financial expenses of moving but also the various policy barriers that deter migration, such as the hukou system and working permits in China.

To sum up, a worker living in location j will migrate to location i if and only if location i provides the highest utility among all locations:

$$\frac{V_i \nu_i}{\lambda_{ij}} \geq \frac{V_{i'} \nu_{i'}}{\lambda_{i'j}}, \forall i' \in \mathcal{J}.$$

[7] The quantity of input bundle requirement for the entry cost is identical across all the locations. However, as the price of the input bundle differs across locations in the equilibrium, the entry costs will also differ in value.

Similar to a standard discrete choice model, conditional on $\{V_i\}$ and $\{\lambda_{ij}\}$, the probability of an individual moving from location j to location i is:

$$m_{ij} = \frac{\left(\frac{V_i}{\lambda_{ij}}\right)^{\kappa}}{\sum\limits_{i'=1}^{J}\left(\frac{V_{i'}}{\lambda_{i'j}}\right)^{\kappa}}. \tag{4}$$

From the law of large numbers, m_{ij} is the fraction of individuals living in location j that move to location i. Denoting the initial population in location j as $\bar{L}_j$, the equilibrium population flow from j to i is $m_{ij}\bar{L}_j$, and the equilibrium population in i is thus:

$$L_i = \sum_{j=1}^{J} m_{ij}\bar{L}_j.$$

2.2 Model Solution

Cost Minimization Solving the cost minimization problem of firm k in location j leads to the following expression for the unit cost of the input bundle, which we denote as $c_j(k)$:

$$c_j(k) = (1-\beta_j)^{\beta_j - 1}\,\beta_j^{-\beta_j}\,[(1+\tau_\ell(k))\,w_j]^{\beta_j}\,(P_j)^{1-\beta_j} = (1+\tau_\ell(k))^{\beta_j}\,\bar{c}_j,$$

where

$$\bar{c}_j = (1-\beta_j)^{\beta_j-1}\beta_j^{-\beta_j}w_j^{\beta_j}P_j^{1-\beta_j},$$

is the cost of a frictionless input bundle. In the expression above, the firm perceives the unit cost of labor as $[1+\tau_\ell(k)]\,w_j$. The unit cost of production is thus firm-specific and different from a standard Melitz model. Higher labor friction increases the costs of workers relative to the composite varieties, distorting the optimal input composition of the firm.

Price, Sales, Revenue, and Employment To maximize the utility described in equation (1), we obtain the following demand function for goods k in location i from the firm located in location j:

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

where $q_{ij}(k)$ and $p_{ij}(k)$ are the quantity and price of variety k sold in location i from location j, respectively. The variable P_i is the ideal price index, and X_i is the total expenditure in location i. Solving the maximization problem leads to the standard pricing decision:

$$p_{ij}(k)=\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c_j(k)}{A_j a(k)\left(1-\tau_y(k)\right)},\tag{5}$$

and the firm's profit derived from market i is:

$$\pi_{ij}\left(k\right)=\frac{1}{\varepsilon}\frac{\left(1-\tau_y(k)\right)X_i}{P_i^{1-\varepsilon}}\left[\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c_j(k)}{\left(1-\tau_y(k)\right)A_j a(k)}\right]^{1-\varepsilon}.\tag{6}$$

Firm-level distortion The pre-tax revenue of the firm is the summation of sales to all the destination markets:

$$r_j(k)=\sum_{i=1}^{J}\left[\frac{X_i}{P_i^{1-\varepsilon}}\left[\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c_j(k)}{\left(1-\tau_y(k)\right)A_j a(k)}\right]^{1-\varepsilon}\right],$$

and the output tax (subsidy) generated by the firm is $\tau_y \cdot r_j(k)$. To account for the payroll friction, first note that the employment of the firm, $\ell_j(k)$, is:

$$\ell_j(k)=f_e\frac{\beta_j\bar{c}}{w_j}+b_j(k)\left[\frac{\beta_j c_j(k)}{(1+\tau_\ell(k))w_j}\right],\tag{7}$$

where $b_j(k)$ is the total number of input bundles required by firm k to cover the production costs:

$$b_j(k)=\sum_{i=1}^{J}b_{ij}(k),\quad b_{ij}(k)=t_{ij}\frac{q_{ij}(k)}{A_j a(k)}.$$

The first term in equation (7) is the number of workers employed to cover the entry costs, and the second term summarizes all the workers employed to cover the production costs. The expression in the square bracket is the number of workers employed to produce a single input bundle. The labor tax (subsidy) paid by the firm is $\tau_\ell \cdot (w_j \ell_j(k) - f_e \beta_j \bar{c}_j)$. In this expression, $w_j \ell_j(k)$ is the total payroll of the firm, from which we deduct the payroll at the entry stage, $f_e \beta_j \bar{c}_j$ since the entry costs are not distorted.

Entry Decision We assume infinitely many potential entrants with zero outside options reside in each location. This assumption implies in the equilibrium, the expected profit before entry should equal the entry costs in each location. Specifically, the free entry condition in location j is:

$$f_e \bar{c}_j = \sum_{i=1}^{J} \iiint_{\underline{a}}^{\infty} \pi_{ij}(k) g_j(a, \tau_y, \tau_\ell) da \cdot d\tau_y \cdot d\tau_\ell. \tag{8}$$

In the expression above, the output and labor frictions affect the entry decisions in the general equilibrium through the profit function, $\pi_{ij}(\cdot)$. The equation also implies that the equilibrium aggregate profit is zero in all the locations.

Price Index The total number of varieties available in location i equals the number of firms that sell to location i from all the locations:

$$\sum_{j=1}^{J} I_j \iiint_{\underline{a}}^{\infty} g_j(a, \tau_y, \tau_\ell) da \cdot d\tau_y \cdot d\tau_\ell,$$

where $g_j(a, \tau_y, \tau_\ell)$ is the probability density function (PDF) associated with $G_j(a, \tau_y, \tau_\ell)$. The ideal price index in location i is a CES aggregator over the prices of all the varieties available in location i:

$$(P_i)^{1-\varepsilon} = \sum_{j=1}^{J} I_j \left(\frac{t_{ij}\varepsilon}{\varepsilon - 1} \right)^{1-\varepsilon} \iiint_{\underline{a}}^{\infty} \left[\frac{c_j(k)}{A_j a (1 - \tau_y)} \right]^{1-\varepsilon} g_j(a, \tau_y, \tau_\ell) da \cdot d\tau_y \cdot d\tau_\ell. \tag{9}$$

Labor Markets The labor demand in location j is the summation of the firm-level employment as in equation (7) across all the operating firms:

$$L_j^d = I_j \iiint_{\underline{a}}^{\infty} l_j(a, \tau_y, \tau_\ell) g_j(a, \tau_y, \tau_\ell) da \cdot d\tau_y \cdot d\tau_\ell.$$

The labor market clearing condition equalizes the labor demand with supply, which consists of migration flows from all the locations $i \in \mathcal{J}$:

$$L_j^d = \sum_{i=1}^{J} m_{ji} \bar{L}_i. \tag{10}$$

Income and Expenditure The total expenditure of location j, X_j consists of two parts, 1) the demand for consumption from workers, and 2) the demand from firms in order to produce input bundles:

$$X_j = Y_j + X_j^b. \tag{11}$$

In the equation above, Y_j is the income in location j, which equals the consumption demand in the equilibrium. The second term, X_j^b, is the demand for varieties to produce input bundles. The equilibrium income in location j includes the labor income and the lump-sum transfer/tax led by both distortions.

$$Y_j = w_j L_j + \gamma_j L_j,$$

where the total lump-sum transfer is:

$$\gamma_j L_j = I_j \iiint \left[\tau_y \cdot r_j(a, \tau_y, \tau_l) + \tau_\ell \cdot (w_j \ell_j(a, \tau_y, \tau_l) - f_e \beta_j \bar{c}_j)\right] \cdot g_j(a, \tau_y, \tau_\ell) da \cdot d\tau_y \cdot d\tau_\ell,$$

in which the first term, $\tau_y \cdot r_j(a, \tau_y, \tau_l)$, is the output tax (subsidy), and the second term, the payroll tax (subsidy). Lastly, the expenditure for producing input bundles is

$$X_j^b = (1 - \beta_j)\, I_j \iiint \left[b_j(a, \tau_y, \tau_\ell) c_j(\tau_\ell) + f_e \bar{c}\right] \cdot g_j(a, \tau_y, \tau_\ell) da \cdot d\tau_y \cdot d\tau_\ell.$$

2.3 Equilibrium

Definition: Conditional on parameters, the equilibrium is a series of wage and price $\{w_j, P_j\}_{j=1}^J$, a series of population distribution, $\{L_j\}_{j=1}^J$, a series of mass of entering firms and expenditure, $\{I_j, X_j\}_{j=1}^J$, such that the following conditions hold:

1. Workers maximize utilities by choosing the consumption of each variety and a location to live, as in equation (4).

2. Firms maximize profits by choosing the price and quantity sold in each market as in equations (5) and (6).

3. The free entry condition holds in each location j, so the expected profit from entry is zero, as in equation (8).

4. The labor market clears in each location as described in equation (10).

5. The balance of trade holds in each location: $X_j = \sum_{i=1}^{J} X_{ij}$, where X_j is defined in equation (11), and X_{ij} is the sales from location j to i.

6. The government budget in each location is balanced so that the total deficit (surplus) incurred by the output and labor frictions is offset by the lump-sum tax on (transfer to) the workers in each location.

Appendix B provides more details on solving the model.

3 Estimation of Firm-level Frictions and Productivity

In this section, we briefly describe a procedure to structurally estimate the critical object of interest in the literature, the joint distribution of the firm-level heterogeneity, $G_j(a, \tau_y, \tau_\ell)$, in each location. While $G_j(a, \tau_y, \tau_\ell)$ is the distribution that all entrants draw their frictions and productivity, what we observe in the data is a subset of entrants who draw favorable combinations of the shocks. Estimation without accounting for selection leads to biased results, as noted in Yang (2021) and Bai et al. (2019). The solution often relies on structurally estimating the distribution by fully solving the general equilibrium. However, doing so is computationally costly, even with a single location, as seen in Bai et al. (2019). Structural estimation by fully solving the general equilibrium is prohibitively expensive in our context with many locations. To overcome this difficulty, we propose an estimation procedure that solves the selection issue by sampling and relies on moment conditions that can be approximated without solving the general equilibrium.

3.1 Data

ASIF The firm-level panel data come from the *Annual Surveys of Industrial Firms* (ASIF) conducted by the National Bureau of Statistics (NBS) of China from 1998 to 2007. The survey covers all state-owned and private firms with more than 5 million RMB in annual sales. We locate the firms by zip codes and restrict our analysis to the prefectures with at least 500 firms in the ASIF. The restriction leaves 237 prefectures in our sample, as depicted in Figure A.1 in the appendix. Our sample is representative, covering more than 74.7 percent of the national GDP in 2007. The number of firms in these prefectures comprises approximately 98.8 percent of the entire sample in the ASIF. We use the information on the sales, payroll, and total costs of production of each firm in the estimation procedure described below.

3.2 Basic Assumptions

To estimate the joint distribution, we assume that the marginal distributions of $1 - \tau_y$ and $1 + \tau_\ell$ are log-normal:

$$\log(1 - \tau_y) \sim \mathcal{N}(0, \sigma_{y,j})$$
$$\log(1 + \tau_\ell) \sim \mathcal{N}(0, \sigma_{\ell,j}),$$

and the marginal distribution of a is a Type-I Pareto with the following CDF:

$$G_{aj}(a) = 1 - a^{-\theta_j}.$$

The three shocks are also correlated with Kendall's rank correlation matrix Σ_j, in which the pair-wise correlations of interests are $\rho_{ay,j}$, $\rho_{a\ell,j}$, and $\rho_{y\ell,j}$.[8] These assumptions imply that we need to estimate six parameters in each location summarized in vector $\delta_j = \{\sigma_{y,j}, \sigma_{\ell,j}, \theta_j, \rho_{ay,j}, \rho_{a\ell,j}, \rho_{y\ell,j}\}$. In total, there are $6J$ parameters across all the locations.

The estimation procedure is designed to recover the distribution of frictions and produc-

[8] We specify the correlation as Kendall's correlation instead of Pearson's linear correlation because Kendall's correlation is preserved under monotonic transformations later used in the copula functions.

tivity within each location. First, note that the above setup assumes that within each location, the marginal distributions of $\log(1-\tau_y)$ and $\log(1+\tau_\ell)$ have zero means. Therefore the implied distortion directly leads to within-location misallocation.[9] Second, as discussed later, we estimate these distributions separately for each prefecture without directly accounting for the cross-location misallocation in the data. However, in the model, the within-location distortions enable the reallocation of firms and workers across space through general equilibrium effects on output and factor prices. Therefore, within-prefecture frictions generate spillovers across space in the model. We conduct robustness checks by relaxing the "zero-mean" assumption and report the results in Appendix D. The counterfactual results reveal interesting and subtle interactions between selection bias and friction distribution. We refer the readers to the appendix for a detailed discussion of these results.

The critical parameters of interest are $\sigma_{y,j}$ and $\sigma_{\ell,j}$, which measure the dispersion of frictions within a location. Higher dispersion implies that micro-frictions are more prevalent within a prefecture, leading to higher economic costs. Similarly, correlations between friction and productivity, $\rho_{ay,j}$ or $\rho_{a\ell,j}$, which we denote as the "productivity correlation" of friction, can also be costly. A positive correlation implies that the more productive firms are taxed more, leading to potentially inefficient resource allocation. Similarly, a negative correlation implies that the more productive firms are favored at the expense of the smaller ones. The consequence of the negative correlation at the aggregate level, however, is more complicated. On the one hand, the negative correlation implies that small firms are more inefficient, discouraging entry and reducing the number of varieties available to consumers, thus leading to welfare losses. On the other hand, however, the negative correlation also implies that more productive firms face lower friction, which might boost overall output and aggregate welfare. Later in the quantitative section, we show that the welfare impact of negative correlations is ambiguous due to the confounding impacts of the two forces described above.

We assume that the marginal distribution of productivity is Pareto, different from Bai et al. (2019), which assumes multivariate normal distribution for both frictions and productivity. The Pareto assumption facilitates the aggregation as many analytical results from Melitz

[9] From the properties of log-normal distribution, the mean of τ_y is $1-\exp(\sigma_{y,j}^2/2)$ and the mean of τ_ℓ is $\exp(\sigma_{\ell,j}^2/2)-1$. Since $\sigma_{y,j}$ and $\sigma_{\ell,j}$ are location-specific, the average friction in a location is mechanically tied to the estimated dispersion parameter.

(2003) depend crucially on it. The tractability at the aggregate level removes the need to approximate the productivity shocks numerically and thus reduces the computational load and improves accuracy when solving the model. Moving away from Gaussian assumptions, however, complicates the estimation strategy. With mixed types of marginal distributions, the joint-distribution $G_j(\cdot)$ described above does not adopt an explicit functional form. As a result, we cannot use Maximum Likelihood Estimators (MLE); instead, we rely on the Simulated Methods of Moments (SMM) and copula functions to recover δ_j. Moreover, the absence of a multivariate Gaussian structure also leads to complexities in estimating the conditional distribution of productivity. As explained later, we resort to Monte-Carlo simulations to circumvent this issue.

3.3 Simulated Method of Moments

Simulation and Sample Selection The first step to implement the SMM is to draw (a, τ_y, τ_ℓ) from the joint distribution specified above, conditional on a parameter vector δ_j. We use copula functions to simulate the firm-level shocks. To do so, we define a Gaussian copula over the interval $[0,1]^3$, which can specifically be expressed as $C^R(u) = \Psi_R(\Psi^{-1}(u_1), \Psi^{-1}(u_2), \Psi^{-1}(u_3))$. Ψ^{-1} is the inverse cumulative distribution function of a standard normal distribution. Ψ_R is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and Pearson's correlation matrix R, converted from Kendall's correlation matrix Σ_j. We can then use the Gaussian copula to generate three dependent random variables valued between 0 and 1. Finally, we apply the inverse CDF methods to convert them into (a, τ_y, τ_ℓ) with the marginal distributions parameterized by δ_j. Within a prefecture, we draw N_j firms from $G_j(\cdot)$ by repeating this process.

We address the selection bias by sample selection at the simulation stage. Selection bias in the data arises because ASIF only includes firms above a revenue threshold. To account for this, we rank the simulated firms by revenue and only compute moment conditions using the largest $\widehat{N}_j < N_j$ firms in each prefecture. In practice, we set N_j to be the number of firms observed in the 2008 Economic Census, which approximates the total number of firms in the prefecture j and set $\widehat{N}_j$ to be the number of firms in ASIF from the same prefecture. The rationale of our procedure is intuitive: prefectures with a large gap between the observed $\widehat{N}_j$ and N_j must be hard for firms to operate in. Drawing a large number of simulated firms and only selecting the top $\widehat{N}_j$ ensures that the simulated firms face fierce selection in our simulation, too.

Conditioning on the observed level of selection addresses the selection bias because the impacts of general equilibrium forces are summarized in the fraction of firms that are above a certain threshold. Counterfactually, if some shock hits the economy that leads to a different distribution of factor prices, these changes would have been reflected in the number of firms covered in ASIF and the economic census. Subsequently, our sample selection procedure would take that shock into account by drawing a different number of N_j and selecting a different $\hat{N}_j$. In other words, conditioning on the level of selection essentially assumes that the observed economy is in equilibrium, and our identified joint distribution is conditional on the distribution of endogenous variables in the observed economy.

The key advantage of selecting by revenue is that the firms within a prefecture can be ranked without knowing any endogenous variables. Put differently, the firm-specific shocks, (a, τ_y, τ_ℓ), are sufficient statistics to rank the firms by revenue. To see this, note that the revenue generated by a firm originating in prefecture j and selling to prefecture i with a draw of (a, τ_y, τ_ℓ) is:

$$r_{ij}(a, \tau_y, \tau_\ell) = \left(\frac{\varepsilon}{\varepsilon - 1}\frac{t_{ij}\bar{c}_j}{A_j}\right)^{1-\varepsilon}\left(\frac{X_i}{P_i^{1-\varepsilon}}\right)\left[a^{\varepsilon-1}\left(1-\tau_y\right)^{\varepsilon}\left(1+\tau_\ell\right)^{\beta_j(1-\varepsilon)}\right],$$

in which the terms in the first two parentheses are common across all the firms in the same location. Therefore, the rank of the market-specific revenue is determined solely by the draws of productivity and friction in the last square bracket. To simplify the notation, we define the terms in the square bracket as the "augmented productivity" inclusive of the frictions, denoted as $\tilde{a}$:

$$\tilde{a}^{\varepsilon-1} = a^{\varepsilon-1}\left(1-\tau_y\right)^{\varepsilon}\left(1+\tau_\ell\right)^{\beta(1-\varepsilon)}.$$

A firm with higher $\tilde{a}$ is more productive, thus selling more to any given market and therefore collecting a higher total revenue. In a random draw of N_j entrants, the top $\widehat{N}_j$ firms ranked by revenue must also be the top $\widehat{N}_j$ firms ranked by $\tilde{a}$. Therefore, the sample-selection procedure based on $\tilde{a}$ addresses the selection bias without solving the general equilibrium.

Moments With the selection bias addressed, we now describe the moment conditions. We structurally estimate the vector δ_j in each prefecture by matching the following 17 moments related to the distribution of productivity and frictions, computed following the methods in Hsieh and Klenow (2009). All the moments in the simulation are computed based on the sample of $\widehat{N}_j$ firms ranked by $\tilde{a}$ and rely on the draws of $\{a, \tau_y, \tau_\ell\}$ without solving the model as emphasized earlier.[10]

The first set of moments are the adjacent percentile ratios of the revenue distribution: the 90-to-75, 75-to-50, 50-to-25, and 25-to-10 ratios. The moments related to the revenue distribution help to pin down θ_j, the dispersion of productivity within a location. In the simulation, the percentile ratios of revenue are the same as those of augmented productivity ($\tilde{a}$).

To pin down the parameters related to τ_y and τ_ℓ, we first follow the approach in Hsieh and Klenow (2009) to estimate firm-level frictions. Solving individual firm k's cost minimization problem leads to the following estimate for $\tau_\ell(k)$:

$$\tau_{\ell,j}(k) = \frac{\beta_j}{1-\beta_j} \cdot \frac{P_j Y_j(k)}{w_j L_j(k)} - 1,$$

where β_j is the location-specific labor intensity that we will discuss later. $P_j Y_j(k)$ is the firm's expenditure on intermediate goods and $w_j L_j(k)$ is the total wage bill. Both variables come from the ASIF. Similarly, the profit maximization problem leads to the estimation of output friction, $\tau_{y,j}(k)$, as follows:

$$\tau_{y,j}(k) = 1 - \frac{1}{1-\beta_j} \frac{\varepsilon}{\varepsilon - 1} \frac{P_j Y_j(k)}{R_j(k)},$$

where $R_j(k)$ denotes firm k's sales revenue, which is also available in the ASIF data.

[10] The estimation also relies on the value of ε, and we follow the literature and set it to 6. Section 4 provides more details regarding the choice of this value.

Our remaining moment conditions are based on the estimated firm-level friction within each prefecture. In particular, we target the standard deviation and the 90-to-75, 75-to-50, 50-to-25, and 25-to-10 percentile ratios for both the distributions of $\log(1-\tau_y)$ and $\log(1+\tau_\ell)$. We have also computed the correlation matrix among $\log(1-\tau_y)$, $\log(1+\tau_\ell)$ and revenue, which serve as the targets for pinning down the parameters governing correlations between frictions and productivity.

We allow β_j to be location-specific in the estimation. One concern regarding the identification of τ_ℓ is that differences in industrial composition across prefectures might be misconstrued as variations of labor friction.[11] To address this concern, we account for the differences in the industrial composition across prefectures in β_j. We compute the labor share of each of the 491 industries in ASIF at the 4-digit classification and then obtain the labor share of each prefecture using the industry composition measured by employment share in each prefecture.[12]

To sum up, we target 17 moments in each prefecture as summarized in the $\mathbf{M}_j$ vector. We denote the simulated moments as $\widehat{\mathbf{M}}_j$. The SMM estimates δ_j by minimizing the weighted distances between $\mathbf{M}_j$ and $\widehat{\mathbf{M}}_j$:

$$\hat{\delta}_j \equiv \operatorname{argmin} \left(\widehat{\mathbf{M}}_j - \mathbf{M}_j\right)^{\mathbf{T}} \mathbf{W}_j \left(\widehat{\mathbf{M}}_j - \mathbf{M}_j\right), \tag{12}$$

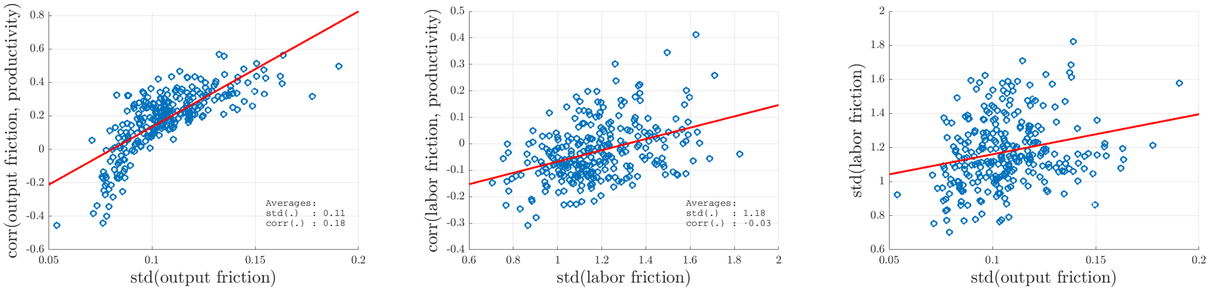
where $\mathbf{W}_j$ is a 17×17 optimal weighting matrix. The weighting matrix is the inverse of the variance-covariance matrix of the data moments computed by bootstrapping the firms within prefecture j. Given the estimated δ_j vector at each prefecture, Figure A.2 in the appendix plots the kernel density estimates of the sales, τ_y, and τ_l distributions in both the data and the simulation for several selected prefectures. Even with the limited number of targeted moments, the simulated distributions closely mimic those in the data.

[11] For example, a prefecture with predominately labor-intensive industries should have higher β_j and thus a higher share of payment to labor in the data. If we measure the firm-level τ_ℓ against a common β, we will incorrectly infer that the firms in the prefecture have low levels of τ_ℓ.

[12] We compute the vector of $\{\beta_j\}$ using the procedure described above outside of the SMM, and therefore β_j is not jointly estimated with the other parameters in δ_j.

3.4 Estimation Results

In this part, we briefly discuss the estimation results. We focus on the standard deviations of the frictions and their correlation with productivity. Note that $\{\sigma_{y,j}, \sigma_{\ell,j}\}$ are the standard deviations of $\log(1 - \tau_y)$ and $\log(1 + \tau_\ell)$, respectively, not the standard deviations of τ_y and τ_ℓ. Similarly, ρ_{ay} and $\rho_{a\ell}$ are the correlations between $\log(1 - \tau_y)$ and $\log(1 + \tau_\ell)$ and a, not the productivity correlations of τ_y and τ_ℓ. To avoid any confusion, we use $\tilde{\sigma}_y$ and $\tilde{\sigma}_\ell$ to denote the standard deviations of τ_y and τ_ℓ, and use $\tilde{\rho}_{ay}$ and $\tilde{\rho}_{a\ell}$ to denote the productivity correlations of τ_y and τ_ℓ. To compute $\{\tilde{\sigma}_y, \tilde{\sigma}_\ell, \tilde{\rho}_{ay}, \tilde{\rho}_{a\ell}\}$, we simulate 100,000 draws of $\{\log(1 - \tau_y), \log(1 + \tau_\ell), a\}$ based on the estimated δ_j in each prefecture, and then use the sample standard deviation of τ_y and τ_ℓ to estimate $\{\tilde{\sigma}_y, \tilde{\sigma}_\ell\}$. Similarly, we use the sample correlations between τ_y, τ_ℓ, and a to estimate $\{\tilde{\rho}_{ay}, \tilde{\rho}_{a\ell}\}$. In this part, we highlight two main data patterns and refer the readers to Appendix D for more details.



(a) Output Frictions

(b) Labor Frictions

(c) Both Frictions

Figure 1: Standard Deviation of Frictions and Correlation with Productivity

Notes: the first panel plots the standard deviations of output frictions, $\tilde{\sigma}_y$, against their correlations with productivity, $\tilde{\rho}_{ay}$. The second panel plots the standard deviations of labor frictions, $\tilde{\sigma}_\ell$, against their correlations with productivity, $\tilde{\rho}_{a\ell}$. The third panel plots the standard deviations of output frictions against those of the labor frictions. Each point in the figures is a prefecture.

The first data pattern is that while both output and labor frictions are prevalent in all the prefectures, the labor frictions are more pronounced. Panel (a) of Figure 1 plots the standard deviation of the output frictions, $\tilde{\sigma}_y$ against their correlation with productivity, $\tilde{\rho}_{ay}$, and the second panel plots the labor frictions similarly. The average standard deviation of output frictions is 0.11 across all 237 prefectures, and that of labor frictions is one order of magnitude higher at 1.18. The large standard deviations in the labor wedges indicate that the input markets are more distorted than the output markets. The correlations with productivity are comparable in size but different in sign. The average correlation between productivity and output friction is positive at 0.18 and with labor friction, slightly negative at -0.03. Moreover, the locations with higher dispersion of the frictions also tend to see higher productivity correlation, and locations with higher dispersion of output friction also tend to see higher dispersion of labor friction. These patterns suggest that the welfare impacts of both frictions could be substantial at both the aggregate and the local levels.

The second pattern is that the locations with higher per-capita GDP tend to experience less dispersion and lower productivity correlation. Panel (a) in Table 1 reports the regressions of both frictions' dispersion and productivity correlation against prefecture-level characteristics. We find that prefectures with higher per-capita GDP enjoy substantially lower dispersion and productivity correlation, controlling total population, industry composition, and connectivity on the transportation networks. The negative coefficients on per-capita GDP are significant in all cases except for $\tilde{\sigma}_{\ell}$. The lower panel of the same table shows that around 13% of the observed spatial variations in per-capita GDP can be attributed to the estimated dispersion and productivity correlation parameters. The explanatory power of output frictions is substantially higher than that of labor frictions.

Figure 2 further highlights these patterns by comparing these measures across four broad regions sorted by per-capita GDP.[13] For example, as seen in Panel (a), while the median dispersion of the output frictions is 0.086 and 0.084 for the two richer regions, the same statistics are 0.108 for the two poorer regions. The correlations between output frictions and productivity are closer to zero in the two richer regions, as shown in Panel (b). While the median correlations are -0.020 and -0.068 on the Southern and the Eastern Coasts, respectively, they are much higher at 0.242 and 0.239 in the two poorer regions. We do not observe significant variations in the spread of labor frictions across regions from Panel (c), similar to the regression results as reported in Table 1. The correlation between labor friction and productivity is closer to zero or negative in all four regions.[14]

[13] See Table A.1 for the definition of regions. The "Southern Coast" is the region with the highest per-capita GDP, followed by the "Eastern Coast". The two relatively less developed regions are the "North"and the "Others" regions.

[14] Table A.2 in the Appendix reports the regression results using regional dummy variables and shows that the regional variations in the dispersion and productivity correlation are statistically significant.

Table 1: Spatial Distribution of Frictions

(a) The Covariates of Frictions

	(1) $\tilde{\sigma}_y$	(2) $\tilde{\sigma}_\ell$	(3) $\tilde{\rho}_{a,y}$	(4) $\tilde{\rho}_{a,\ell}$
log(per-capita GDP)	-0.006**	-0.024	-0.107***	-0.028**
	(0.002)	(0.024)	(0.025)	(0.013)
log(population)	-0.003*	-0.012	-0.034*	0.011
	(0.002)	(0.025)	(0.020)	(0.014)
industry share	0.000	0.002	0.002	0.002***
	(0.000)	(0.001)	(0.002)	(0.001)
labor share	0.254***	-2.864**	1.095	-0.168
	(0.080)	(1.124)	(0.816)	(0.477)
remoteness	0.090***	0.341	0.251	-0.391***
	(0.020)	(0.231)	(0.164)	(0.123)
N	237	237	237	237
Adj.R-squared	0.158	0.023	0.112	0.098

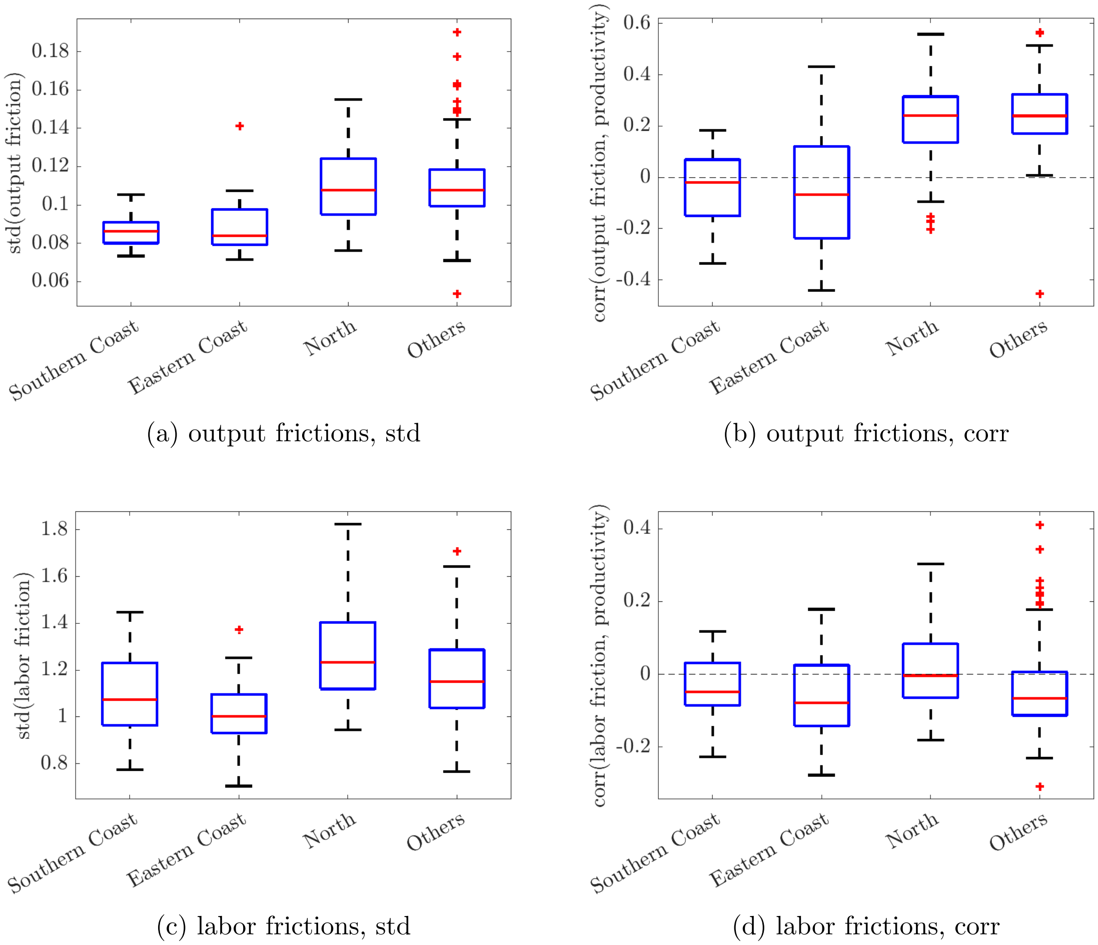
(b) Variations of Per-capita GDP Explained by Frictions

Parameters	R^2	Parameters	R^2
$\tilde{\sigma}_y$	0.0197	$\tilde{\sigma}_y$, $\tilde{\sigma}_\ell$	0.0219
$\tilde{\sigma}_\ell$	0.0000	$\tilde{\rho}_{ay}$, $\tilde{\rho}_{a\ell}$	0.0998
$\tilde{\rho}_y$	0.0913	$\tilde{\sigma}_y$, $\tilde{\rho}_{ay}$	0.1137
$\tilde{\rho}_\ell$	0.0054	$\tilde{\sigma}_\ell, \tilde{\rho}_{a\ell}$	0.0072
$\tilde{\sigma}_y$, $\tilde{\rho}_{ay}, \tilde{\sigma}_\ell$, $\tilde{\rho}_{a\ell}$	0.1332		

Notes: the upper panel reports the regression results of the dispersion and productivity correlation of output and labor frictions against prefecture-level characteristics. Robust standard errors are reported in the parenthesis. ***: significant at the 1% level; *: significant at the 5% level; *: significant at the 10% level. "Industry share" refers to the share of the manufacturing industry in GDP, "labor share" refers to β_j discussed earlier, and "remoteness" of a prefecture measures the location of a prefecture in the transportation network, from Ma and Tang (2020). The lower panel reports the r-squared obtained from regressing log(per-capita GDP) against various combinations of the dispersion and productivity correlation parameters.

3.5 Discretization and the Conditional Distributions of Productivity

Before evaluating the model quantitatively, we discuss our strategy to discretize the joint distribution of friction and productivity. In this part, we propose a novel numerical strategy that takes advantage of the analytical tractability of Pareto-distributed productivity.



(a) output frictions, std

(b) output frictions, corr

(c) labor frictions, std

(d) labor frictions, corr

Figure 2: Frictions by Region

Notes: the four panels display the box plots of the standard deviations of the output and labor frictions, $\tilde{\sigma}_y$ and $\tilde{\sigma}_\ell$, as well as their correlations with productivity, $\tilde{\rho}_{ay}$ and $\tilde{\rho}_{a\ell}$, by regions. The four regions are sorted by average per-capita GDP in descending order. "Southern Coast" refers to the prefectures in the provinces of Guangdong and Hainan; "Eastern Coast" refers to Shanghai and the prefectures in the provinces of Jiangsu, Zhejiang, and Fujian; "North" refers to Beijing, Tianjin, and the prefectures in the provinces of Hebei, Shandong, Liaoning, Jilin, and Heilongjiang. The prefectures in all the other provinces are included in "Others". In the box plot, the central red line marks the median, and the bottom and top edges of the box mark the 25th and the 75th percentiles, respectively. The two bars are the upper and lower adjacent values, defined as $U = x_{[75]} + \frac{3}{2}\left(x_{[75]} - x_{[25]}\right)$ and $L = x_{[75]} - \frac{3}{2}\left(x_{[75]} - x_{[25]}\right)$, where $x_{[25]}$ and $x_{[75]}$ are the 25th and the 75th percentiles of the data. The outliers, marked with the "+" sign, are the observations that are higher than U or lower than L.

To evaluate the integrals that involve firm-level shocks, such as in equation (8) and (9), the traditional approach is to discretize all three dimensions of the joint distribution into grids and then evaluate the integrals numerically. The main drawback of this approach is that the numerical accuracy is highly dependent on the productivity grid. This high sensitivity of productivity comes from the fact that the productivity distribution is highly skewed.[15] In our context, a productivity grid with thousands of points is required to achieve a reasonable level of accuracy. Unfortunately, given the complexity of the model, such a productivity grid is prohibitively expensive to implement. To circumvent this issue, we propose a new approach that analytically evaluates the integration associated with productivity.

Our approach works as follows. We first discretize τ_y and τ_ℓ into an $N_y \times N_\ell$ matrix, in which (τ_y^k, τ_ℓ^l) is the (k, l)th element of the matrix. Conditional on a pair of (τ_y^k, τ_ℓ^l), we estimate the distribution of productivity, assuming that the conditional distribution is Pareto. With the estimated conditional productivity, we then analytically evaluate all the integrals that involve productivity. Compared to the alternative approach of discretizing productivity, our approach is analytically tractable, significantly reducing computational load and improving simulation accuracy.

To implement this solution strategy, we first need to estimate the distribution of productivity, conditional on the realization of τ_y^k and τ_ℓ^l.[16] The estimation is not trivial because the conditional distribution does not adopt a closed-form density function, as is common in Gaussian copulas. The lack of a closed-form density function leads to several issues. The first conceptual issue is that one cannot ascertain the conditional distribution to be Pareto. Fortunately, we find that the assumption of Pareto distribution approximates the conditional distribution of productivity reasonably well. To measure the goodness-of-fit of the Pareto assumption, we compute the adjusted R^2 in the Zipf plots for each (τ_y^k, τ_ℓ^l) in each prefecture j, as detailed later. Out of the $J \times N_y \times N_\ell$ tests, the median adjusted R^2 is 0.999, and 95 percent of the adjusted R^2s are higher than 0.992. Our simulations' goodness-of-fit is better than the commonly-accepted empirical Pareto distributions. For example, Axtell (2001) shows that the firm size distribution in the U.S. follows a Pareto distribution, and the adjusted R^2 of the firm size distribution in the U.S. data is 0.992 when measured using employment, and 0.976 when measured using sales. In the case of the U.S. city-size distribution, the adjusted R^2 of the Zipf plot is 0.986 in Gabaix (1999). Based on this, we assume

[15] In the extended model with fixed costs, the accuracy issue is even worse because the limits of the integration are also endogenous.

[16] One cannot directly use the marginal distribution of productivity as estimated above because the productivity and the frictions are correlated.

the conditional distribution is Pareto and proceed as follows.

Given that the conditional distribution of productivity follows a Pareto distribution, our next step is to estimate the shape parameter, θ^{kl} for each grid point (τ_y^k, τ_ℓ^l).[17] We proceed with Monte-Carlo simulations in the following steps in each prefecture j:

Step 1: Simulate N_{MC} draws of (a, τ_y, τ_ℓ) from the distribution defined in the estimated $\hat{\delta}_j$.

Step 2: Assign each draw, (a, τ_y, τ_ℓ), to the nearest grid point (k, l) in the $N_y \times N_\ell$ matrix. We define the nearest point as

$$k = \text{argmin}_{k'=1,2,\cdots,N_y}|\tau_y - \tau_y^{k'}|, \quad l = \text{argmin}_{l'=1,2,\cdots,N_\ell}|\tau_\ell - \tau_\ell^{l'}|.$$

Step 3: Denote the number of draws assigned to each grid point (k, l) as n^{kl}. Estimate θ^{kl} using the method-of-moments estimator based on n^{kl} draws as:

$$\widehat{\theta}^{kl} = \frac{\bar{a}^{kl}}{\bar{a}^{kl} - \underline{a}^{kl}}, \tag{13}$$

where $\bar{a}^{kl}$ is the *mean* of, and $\underline{a}^{kl}$ is the *minimum* of, all the productivity draws assigned to the grid point. Firm size follows a power law distribution in our model, and the tail index of firms at grid point (k, l) is $\theta^{kl}/(\varepsilon - 1)$. To ensure the existence of the mean, we assume that the tail index is greater than 1, which implies a lower bound of $\theta^{kl} > \varepsilon - 1$ in the estimation.

Step 4: To measure the goodness-of-fit of the Pareto assumption, we compute the adjusted R^2 of the Zipf plot at each grid point (k, l). The exercise plots the logarithm of a against the logarithm of the frequency of a at each grid point (k, l). If the underlying distribution is Pareto, Zipf's plot would be linear. With this insight, we run a simple linear regression on the log-log plot and use the adjusted R^2 to measure the goodness-of-fit of the Pareto assumption at the grid point.

[17] We have also estimated the location parameter of the Pareto distribution for each grid point and find that they are all clustered at 1, the location parameter in the marginal distribution. As a result, we assume the location parameter in all the conditional distributions to be 1.

In practice, we set $N_y = N_\ell = 5$ and $N_{\text{MC}} = 20$ million. We discretize both frictions by standard errors so that the five grid points correspond to $[-2, -1, 0, +1, +2]$ standard deviations away from the mean, respectively. Appendix C provides more details on the estimation, as well as the solution of the model based on the discretized frictions.

4 Calibration

Other than the estimated joint distribution, the remaining parameters to be disciplined are $\{\psi, \kappa, \varepsilon, \beta_j, f_e\}$, the initial population distribution $\{\bar{L}_j\}$, the fundamental productivity and amenity, $\{A_j, \phi_j\}$, trade costs matrix $\{t_{ij}\}$, and the migration cost matrix, $\{\lambda_{ij}\}$. We calibrate the model into the same 237 prefectures as defined in the previous section from 1998 to 2007. Table 2 summarizes the calibrated parameters.

The following parameters are calibrated without solving the model. The congestion elasticity, $\psi = -0.1$, comes from Ahlfeldt et al. (2015).[18] The migration elasticity, $\kappa = 2.02$, comes from Caliendo et al. (2019). The elasticity of substitution, $\varepsilon = 6$, is consistent with the range commonly used in the literature.[19] We calibrate the location-specific labor share, β_j, by combining the prefecture-specific labor share from ASIF and the national input-output tables. We do not directly use the labor share from ASIF because the dataset only contains large firms and thus does not represent all the firms in China. In particular, we take the prefecture-specific labor share from the ASIF—the same as those estimated in Section 3.3—and normalize them so that the average β_j is 0.37, which is the aggregate labor share derived from the 2002 Input-Output Table of China.[20] This normalization ensures that β_j varies across prefectures as indicated by ASIF and is consistent with the labor share in the national IO tables. Lastly, the initial population distribution $\{\bar{L}_j\}$ comes from the *Population Census* in 2000.

[18] Note that with the absence of agglomeration elasticity, $\psi = -0.1$ ensures the existence and the uniqueness of the equilibrium, as discussed in Allen and Arkolakis (2014).

[19] Anderson and van Wincoop (2004) surveyed the estimates of the elasticity of substitution in the literature and concluded that the reasonable range is between 5 and 10.

[20] Denote the labor share estimated in Section 3.3 at location j as β_j^{SMM}. The normalized labor share used in the quantitative part is $\beta_j = \beta_j^{\text{SMM}} - (1/J) \sum_{i=1}^{J} \beta_i^{\text{SMM}} + 0.37$.

Table 2: Parameters

(a) Pre-set Parameters

Para.	Value	Source	Note
ψ	-0.1	Ahlfeldt et al. (2015)	congestion elasticity
κ	2.02	Caliendo et al. (2019)	migration elasticity
ε	6.0	Anderson and van Wincoop (2004)	elasticity of substitution
$\{\beta_j\}$	-	ASIF data and IO Table	prefecture-specific labor share
$\{\bar{L}_j\}$	-	population census, 2000	prefecture-specific initial population

(b) Jointly-Calibrated Parameters

Para.	Value	Calibration target	Note
f_e	28.5	firms-to-population ratio	entry costs
$\bar{t}$	2.06	internal-trade-to-GDP	average trade costs
$\bar{\lambda}$	39.9	aggregate stay-rate	average migration costs
$\{A_j\}$	-	output	prefecture-specific productivity
$\{\phi_j\}$	-	population	prefecture-specific amenity

Notes: Panel (a) presents the parameter externally determined, and Panel (b) presents the jointly calibrated parameters. ψ comes from Ahlfeldt et al. (2015). The value of κ is from Caliendo et al. (2019), and the value of ε comes from Anderson and van Wincoop (2004). β_j is computed using the ASIF data and Input-Output Tables in China. f_e is calibrated to match the firm-to-population ratio from the *2008 Economic Census*. $\bar{t}$ is calibrated to match the target obtained from the *Investment Climate Survey*. The target for $\bar{\lambda}$ came from the *One Percent Population Survey* in 2005. $\{A_j\}$ matches the prefecture-level output, and $\{\phi_j\}$ matches the prefecture-level population.

Joint Calibration All the other parameters are jointly calibrated in the general equilibrium. We follow the strategy in Ma and Tang (2020, 2022) by assuming that the iceberg trade and migration costs are functions of the geographic costs matrix, T_{ij}, which are estimated in the same paper. In particular, we assume the following functional form and normalize $t_{ii} = \lambda_{ii} = 1, \forall i$:

$$t_{ij} = \bar{t} \times T_{ij}, \forall i \neq j$$
$$\lambda_{ij} = \bar{\lambda} \times T_{ij}, \forall i \neq j.$$

The T_{ij} matrix from Ma and Tang (2020) describes the relative costs of transportation based on the observed road, railway, and river network between all prefecture pairs. The functional form assumption reduces the calibration of the $\{t_{ij}\}$ and $\{\lambda_{ij}\}$ matrices to two scalars, $\bar{t}$ and

$\bar{\lambda}$.

We normalize the vector of the location-specific productivity, $\{A_j\}$, so that the productivity in Beijing is 1. The remaining 236 elements in $\{A_j\}$ are to be calibrated. Similarly, we normalize ϕ_j in Beijing to be 1 and jointly calibrate the other elements in $\{\phi_j\}$. Together with the three remaining scalars, f_e, $\bar{t}$, and $\bar{\lambda}$, we have a total of $3+236*2 = 475$ parameters to be jointly calibrated.

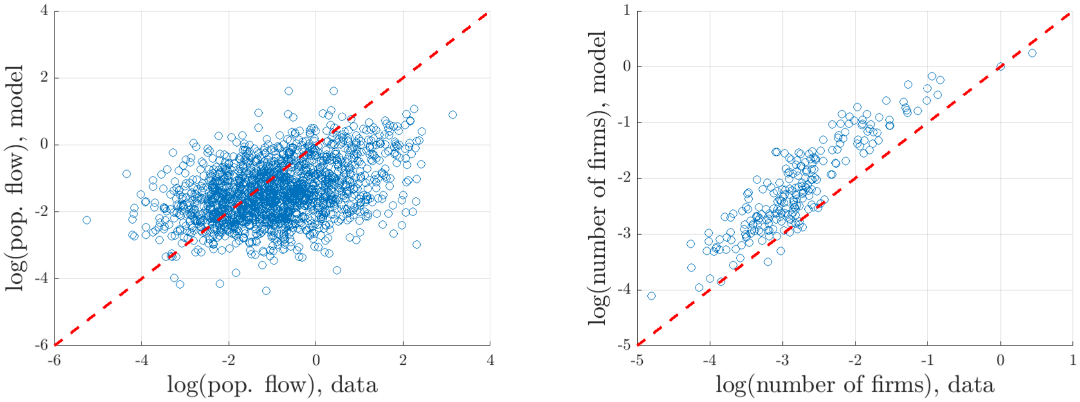
The parameter f_e is the fixed cost of entry. We pin down f_e by matching the ratio of entering firms to the population in the data. In the model, the moment is computed as $\sum_{j=1}^{J} I_j/\bar{L}$. In the data, the number of firms corresponds to the number of legal entities ("Fa Ren") in the *2008 Economic Census*, and the target moment is 5.7 entering firms per thousand population.

The average costs of internal trade, $\bar{t}$, is chosen to match the internal-trade-to-GDP ratio of 0.625 in China, as reported in the *Investment Climate Survey* by the World Bank in 2005. $\bar{\lambda}$ is calibrated to match the aggregate stay rate computed from the *One Percent Population Survey* in 2005. In the data, we define the aggregate stay rate as the fraction of the population that does not move between 2000 and 2005 in the initial population in 2000. We calculate the ratio using a sample of 237 prefectures. In the model, the corresponding statistics is $\sum_{j=1}^{J} m_{jj}\bar{L}_j/\sum_{j=1}^{J}\bar{L}_j$, where m_{jj} is the probability of staying as defined in equation (4).

Lastly, we treat $\{A_j, \phi_j\}$ as the structural residuals and use them to match the real output and population share in each prefecture as reported in the *City Statistical Yearbooks* in 2007.[21] The 475 parameters described above are jointly determined in the general equilibrium and, thus, are jointly calibrated. We implement the joint calibration using a fixed-point algorithm and describe the details in Appendix C.

Model Fit Before discussing the counterfactual results, we present several measures of model fit in Figure 3. The calibration strategy outlined above allows us to perfectly match the real wage and population in each prefecture but not the other aggregate endogenous variables. Nevertheless, our model can capture the salient features in the data, such as migration flows and the number of firms.

[21] We use the summation of each prefecture's secondary and tertiary GDP to proxy its output. The model counterpart is Y_j. We ignore the agriculture output as the model excludes the agriculture sector.



(a) migration flows

(b) number of firms

Figure 3: Model Fit

Notes: the figures plot the model-predicted population flows and the number of firms against the data. The red dashed line is the 45-degree line. In the left panel, each dot is a prefecture pair, and we omitted the pairs where the origin and the destination prefectures are the same. In the right panel, each dot represents a prefecture. The number of firms in the model is I_j. The number of firms in the data comes from the *Economic Census*.

As shown in Figure 3, the bilateral migration flows in our model broadly match that in the data. The ability to match the bilateral migration flows stems from the assumption that the migration costs depend on distance through transportation costs. The reliance on distance echoes the findings in the literature that a gravity-like relationship holds for the international migration flows, such as in Grogger and Hanson (2011) and Ortega and Peri (2014). Panel (b) of the same figure suggests that we can reasonably match the prefecture-level number of firms.

5 Quantitative Results

We evaluate the aggregate and the distributional impacts of micro-level frictions with counterfactual exercises. In the first set of counterfactual simulations, we lower the dispersion of the frictions, $\sigma_{y,j}$ and $\sigma_{\ell,j}$, and in the second set, we lower the productivity correlations, $\rho_{ay,j}$ and $\rho_{a\ell,j}$. To carry out a counterfactual exercise, we repeat the steps outlined in Section 3.5 to re-discretize the output and labor friction grids and re-estimate the conditional distribution of productivity. We then solve the model based on the counterfactual grid points. All the other parameters are the same between the counterfactual and the baseline simulations.

We define the *equilibrium* welfare of a prefecture, $\bar{V}_j$, as:

$$\bar{V}_j \equiv \frac{w_j + \gamma_j}{P_j} \left(L_j\right)^{\psi} \left(m_{jj}\right)^{-\frac{1}{\kappa}}. \tag{14}$$

Compared to indirect utility, V_j, defined in equation (3), the equilibrium welfare, $\bar{V}_j$, also captures the impacts of the migration frictions and idiosyncratic shocks, ν, by including the last term $(m_{jj})^{-\frac{1}{\kappa}}$.[22] The national-level welfare is then computed as $\sum_{j=1}^{J} L_j \bar{V}_j$. In the paper, we focus on the welfare implications and relay those based on the real income to the Appendix Figures A.3 to A.6.[23]

5.1 Dispersion and Correlation of Frictions

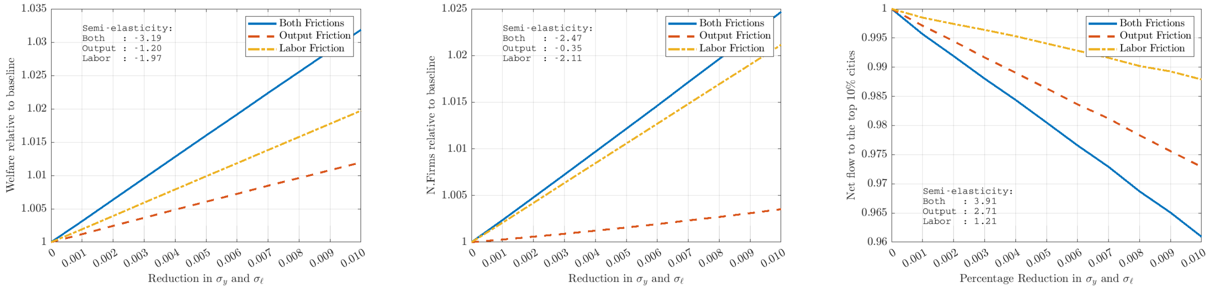
In the first set of results, we reduce the levels of $\{\sigma_{y,j}, \sigma_{\ell,j}\}$ by 0.01 in all the prefectures and compute the semi-elasticity of various endogenous variables to the dispersion parameters.[24] With lowered dispersion of the frictions, misallocation is alleviated in all the prefectures. The impacts at the aggregate level are summarized in Figure 4, and the impacts at the prefecture-level are reported in Figure 5.

The within-prefecture misallocation exerts a sizable impact on the welfare at the national level. As shown in Panel (a) of Figure 4, reducing σ_y and σ_ℓ by 0.01 increases the aggregate welfare by 3.19 percent, leading to a semi-elasticity of -3.19. When the distributions of frictions become less dispersed, the firms are more likely to draw frictions closer to zero within each prefecture, which improves aggregate welfare through several channels. As the labor frictions converge towards zero within the prefecture, the marginal product of labor is more equalized across firms; similarly, the reduction in the dispersion of output frictions better aligns market share with firm-level productivity. Both forces lead to aggregate gains

[22] See proposition 2 in Tombe and Zhu (2019) for more details.

[23] We denote Y_j/P_j as the "real income" of prefecture j, and compute the aggregate real income as $\sum_{j=1}^{J}(Y_j/P_j)L_j$. The real income includes real wage, w_j/P_j, and transfer payments, γ_j/P_j.

[24] A reduction of 0.01 is different from a proportional reduction of 1%. For example, if $\sigma_{y,j} = 0.5$ in the baseline, a reduction of 0.01 means that in the counterfactual $\sigma_{y,j} = 0.49$. We report the percentage reduction results and the implied elasticity in Appendix Figures A.4 and A.5 and discuss these later in the section.



(a) aggregate welfare (b) firm entry (c) net flow to rich prefectures

Figure 4: The Aggregate Impacts of Reducing $\sigma_{y,j}$ and $\sigma_{\ell,j}$

Notes: the panels present the aggregate and distributional impacts of reducing the dispersion of frictions in all the prefectures. The x-axis in all panels indicates the reduction in $\sigma_{y,j}$ and $\sigma_{\ell,j}$ in levels. The y-axes in panels (a), (b), and (c) denote the ratio of aggregate welfare, the number of entering firms, and the net population flow to rich prefectures to the baseline level, respectively.

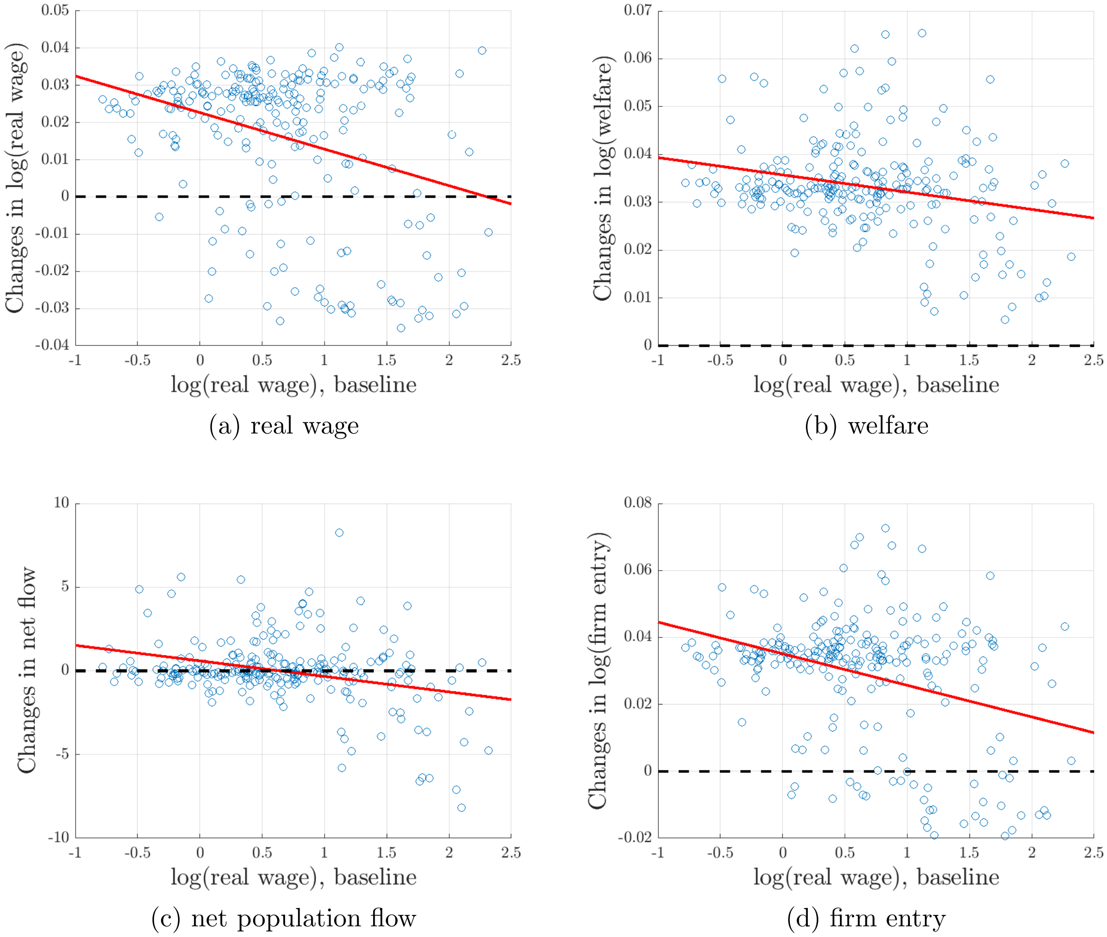
in output. At the extensive margin, the higher expected profit encourages more firms to enter, further lowering the price index, as shown in Panel (b) of the same figure. Lastly, as we will discuss later in Panel (c), the dispersion reduction also reallocates the population towards smaller prefectures, alleviating the congestion disutility.

To disentangle the impacts of each friction, we repeat the exercises by only lowering σ_y and σ_ℓ separately. The semi-elasticity of welfare to labor dispersion (-1.97) is higher than that of the output friction (-1.20), as shown in Panel (a) of Figure 4. The differences between the labor and the output frictions are even more pronounced in the number of entering firms, as shown in Panel (b) of the same figure. The greater impact of labor friction is mainly because the estimated labor dispersion is much larger than the output dispersion, as discussed earlier.

In Panel (c) of Figure 4, we plot the changes in the net population flow to the "top-10 richest prefectures" from the baseline equilibrium against the reductions in the standard deviation.[25] As mentioned earlier, reducing the standard deviation encourages migration toward poorer prefectures. When the standard deviation of both frictions drops by 0.01, the migration flow toward those richest prefectures declines by around 3.91 percent.

At the prefecture level, the changes in local welfare and the real wage reduce population

[25] "Net population flow" in prefecture j is defined as the difference between the equilibrium and the initial population, $L_j - \bar{L}_j$. The "top-10 richest prefecture" is measured by the real wage in the baseline equilibrium.



(a) real wage

(b) welfare

(c) net population flow

(d) firm entry

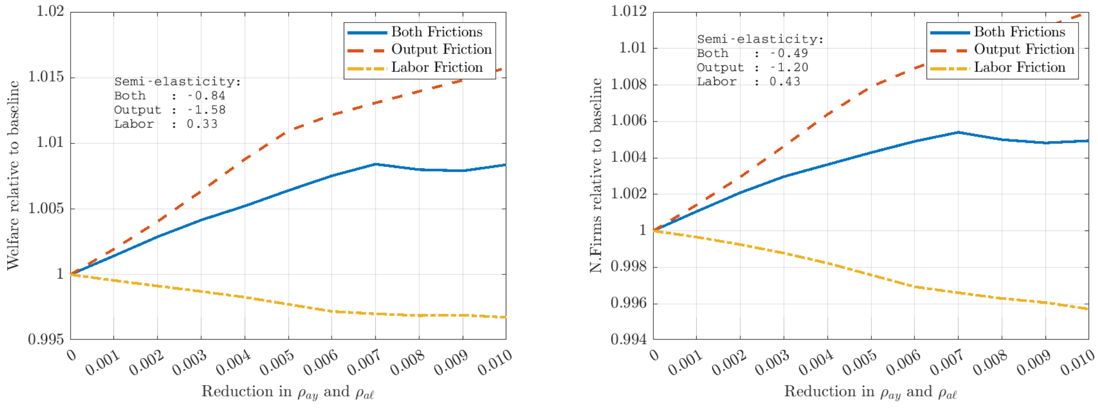
Figure 5: The Prefecture-Level Impacts of Reducing $\sigma_{y,j}$ and $\sigma_{\ell,j}$ by 0.01

Notes: the panels present the distributional impacts of reducing σ_y and σ_ℓ by 0.01. The x-axis in all panels indicates the natural logarithm of the initial real wage. The y-axes in panels (a), (b), and (d) are log differences between the counterfactual and the baseline levels of the real wage, welfare, and the number of entering firms, respectively. Panel (c) plots the differences in the level of net population flow in the unit of 10,000 people. Each dot represents a prefecture.

flows toward richer prefectures. Figure 5 plots the prefecture-level response to a 0.01 reduction in $\{\sigma_{y,j}, \sigma_{\ell,j}\}$. Panels (a) and (b) show a negative relationship between a prefecture's initial real wage and the changes in real wage and welfare. The poorer prefectures benefit disproportionately more from the reduction in the dispersion and thus become more attractive. As a result, economic activity reallocates from the richer prefectures towards the poorer ones, as seen in Panel (c) in both Figures 4 and 5 in the case of population and in Panel (d) of Figure 5 in the case of firm entry. The byproduct of the reallocation is a decline in the spatial inequality of welfare. Our results show that with the 0.01 reduction in both standard deviations, the Gini coefficient of real wage decreases by 0.26%, with a semi-elasticity of -0.26. Figure A.3 in the appendix shows that the results remain qualitatively the same with other common measures of spatial inequality, such as the logarithm of the standard deviation and the Herfindahl-Hirschman index.

The equalizing effect of a uniform reduction in the dispersion suggests a stronger impact in locations with initially higher σ, and the impact weakens in prefectures with initially lower σ. Poorer prefectures face higher dispersion in the benchmark economy as shown in Table 1, and thus benefit more from the uniform decline in the dispersion parameters. The findings suggest "decreasing-returns-to-scale" in reducing the standard deviations of the frictions: a larger proportion of reduction in the standard deviation leads to smaller gain in real wage.

Productivity Correlation In addition to the dispersion parameters, we also study the impact of the productivity correlations, $\{\rho_{ay}, \rho_{a\ell}\}$, by reducing the absolute values of these parameters. When the correlation moves closer to zero, productivity becomes less dependent on friction.



(a) aggregate welfare

(b) firm entry

Figure 6: The Aggregate Impacts of Reducing $\rho_{ay,j}$ and $\rho_{a\ell,j}$

Notes: the panels present the aggregate impacts of reducing the productivity correlation of frictions in all the prefectures. The x-axis indicates the reduction in $\rho_{ay,j}$ and $\rho_{a\ell,j}$ in absolute values. The y-axes are the ratio of the aggregate welfare and the number of entering firms to the baseline levels, respectively.

The reduction of the productivity correlation leads to similar results compared to a decrease in the dispersion. As shown in Figure 6, when the absolute value of the correlation parameters declines by 0.01, the aggregate welfare increases by 0.84 percent. The sign of productivity correlation is crucial for the welfare implications. Intuitively, while reducing a positive correlation makes more productive firms less distorted and thus reallocates resources towards them, moving a negative correlation closer to zero achieves exactly the opposite effect by distorting the more productive firms more. As a result, reducing the absolute value of the negative productivity correlations could dampen welfare. In the baseline estimation, the welfare-dampening impact primarily applies to labor frictions because as many as 146 prefectures have negative correlations between labor frictions and productivity. Subsequently, reducing the absolute values of productivity correlations with labor frictions leads to lower welfare in many prefectures. In comparison, only 38 prefectures have negative correlations between output friction and productivity, and therefore, we observe welfare improvements when correlations between output friction and productivity are lowered.

Elasticities The results discussed above are based on uniformly reducing the key parameters by 0.01 (semi-elasticities). Appendix Figures A.4 and A.5 present the results based on the elasticity of σ and ρ, respectively. To compute the elasticity, we reduce σ-parameters by 10% or the absolute values of the ρ-parameters by 10%. The results are consistent with those reported in Figures 4 and 6. One notable difference is that the elasticity of labor dispersion (-1.61) is much larger than that of the output dispersion (-0.13). The difference comes from the fact that the average standard deviation of the labor friction is substantially higher than that of the output friction, as shown in Figure 1 and discussed in Section 3.4. Due to the differences in the magnitude, the same percentage reduction in dispersion parameters means a much more significant reduction in levels of labor frictions, although their semi-elasticities are similar as shown in Figures 4 and 6.

5.2 Migration and Spatial Frictions

In the last part, we study the interaction between migration friction and the spatial distribution of firm-level frictions. We first compute a "no migration" baseline equilibrium by setting $\bar{\lambda}$ to a sufficiently high value so that the aggregate stay rate increases to 1.0. All the other parameters are the same as in the baseline model. We then reduce the dispersion and productivity correlation parameters starting from the "no migration" baseline. Comparing the effects of reducing friction with and without migration highlights the interaction between the two. Table 3 summarizes the results.

Table 3: The Effects of Migration (Percentage Points)

	Aggregate Welfare			Firm Entry			Gini		
	Baseline	No Mig.	Δ	Baseline	No Mig.	Δ	Baseline	No Mig.	Δ
σ	3.19	3.37	5.83	2.47	2.63	6.59	-0.30	-0.25	-19.28
ρ	0.84	0.91	9.16	0.49	0.61	24.18	-0.08	-0.05	-32.65

Notes: the table summarizes the effects of reducing σ or ρ by 0.01 for both output and labor frictions in all prefectures, with or without internal migration. All the welfare impacts are in percentage points. For example, reducing σ by 0.01 increases aggregate welfare by 3.19% in the baseline model and 3.37% in the model without migration. Therefore, the effect of migration is $3.37/3.19 - 1 = 5.83\%$, as shown in the third column under the header "Δ". The Gini coefficient refers to that of the real wage.

Shutting down the migration *amplifies* the aggregate impacts and *dampens* the distributional impacts of micro-frictions. As shown in Table 3, in the model without migration, reducing σ_y and σ_ℓ by 0.01 increases aggregate welfare by 3.37 percent, higher than the effect of the baseline model with migration at 3.19 percent. Therefore, shutting down migration amplifies the aggregate impacts by $3.37/3.19-1 = 5.83$ percent. Similar results emerge when the productivity correlation is lower, as seen in the second row of the same table. In the baseline model with migrations, the negative impacts of micro-level frictions on aggregate output are partially offset by the migration flows — people leaving the heavily distorted regions in favor of the less distorted ones, thus reducing the economic activity in the heavily distorted regions. Without migration, the aggregate impacts are amplified as people can no longer escape the heavily distorted prefectures.

Moreover, without migration, spatial inequality becomes less responsive to micro-frictions. For example, lowering the dispersion parameters reduces the Gini coefficient by 0.3 percent with migration, but the impact reduces to 0.25 percent when migration is not allowed. The impact on spatial inequality is the by-product of the mechanism described above. Micro-frictions in our model lead to spatial inequality partly because they divert away workers. Without migration, this negative channel is shut down. The heavily distorted regions no longer suffer from an exodus of workers; therefore, the distributional impacts are less severe.

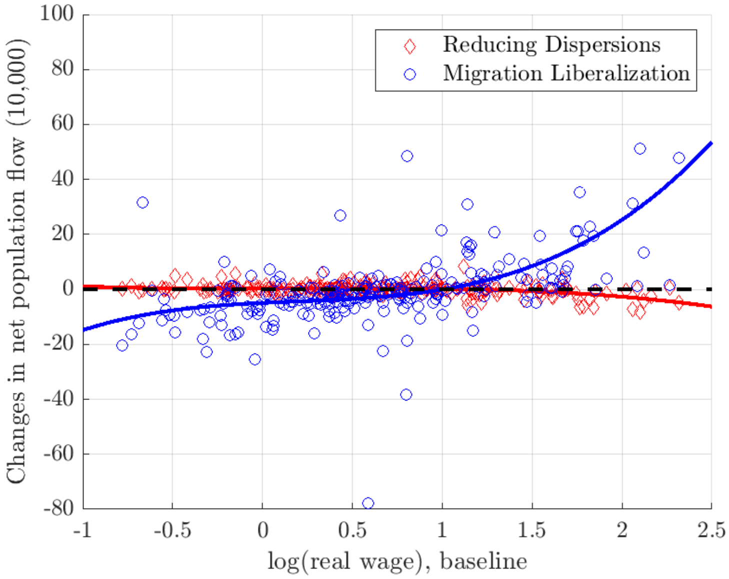


Figure 7: Reducing dispersion parameter v.s. migration frictions

Notes: the red and blue dots (curve) represent the distributional impacts from reducing the standard deviation of frictions and migration liberalization in all the prefectures, respectively. The x-axis is the natural logarithm of the initial real wage. The y-axis is the log differences in the net population flow to receiving prefectures from the baseline to the counterfactual economy. Each dot is a prefecture, and the curves are the best quadratic fit.

The policy implication of the exercises above is that migration and place-based policies that reduce local friction are substitutes for impacting aggregate welfare. In other words, the government can improve aggregate welfare by either moving people away from the less-developed regions or alleviating micro-frictions in these regions. By our calculation, to achieve the same level of welfare gain as reducing the standard deviations of both friction by 0.01, the migration frictions, $\bar{\lambda}$, need to decline by around 14.4 percent from the baseline level. The liberalization in migration frictions increases the total number of migrants by 3.0 percent, leading to the same 3.19 percent increase in aggregate welfare. However, the two policies have drastically different implications for spatial inequality. When the migration barriers are alleviated, people move to the richer prefectures. When the distortions are less dispersed, people move to the poorer prefectures, as seen in Figure 7. As a result, the Gini coefficient of real wage increases with migration liberalization by 0.002 and decreases under less dispersed micro-level distortions by 0.001.

5.3 Robustness Checks

Table 4 summarizes the robustness of the main results. The upper panel reports the semi-elasticity of endogenous variables of interest to measures of dispersion, σ_y and σ_ℓ, and productivity correlations, ρ_{ay} and $\rho_{a\ell}$, and the lower panel reports the effects of migration. We repeat the baseline parameterization in the first row for reference in both panels. In all the robustness checks, we re-calibrate the parameters reported in Panel (b) of Table 2 because the calibration targets are endogenous.

Elasticity Parameters The first three sets of the robustness checks focus on the values of ψ, κ, and ε, which we directly take from the typical estimates in the literature. The semi-elasticities are robust across the various specifications.

Across all the alternative parameters, shutting down migration amplifies the aggregate implications while dampening the distributional effects of the dispersion parameter. The effects of migration are weaker with high congestion elasticity when $\psi = -0.2$ because congestion forces discourage workers from moving into the larger and less-distorted prefectures. Similarly, lower migration elasticity ($\kappa = 1.5$) also weakens the migration effect as workers are less likely to move to locations with higher real income. Lower elasticity of substitution ($\varepsilon = 5$) achieves the opposite results because it strengthens the love-of-variety effect and thus encourages workers to move to richer prefectures that host more varieties.

Fixed Exporting Costs The baseline model assumes away the fixed costs of exporting and production. As a robustness check, we allow such forces following the spirit of Melitz (2003). In this extended model, to "export" from location j to i, a fixed cost of f_{ij} in the unit of input bundle in location j is incurred. Upon entry, the firm draws its productivity and frictions from the joint distribution $G_j(a, \tau_y, \tau_\ell)$. Conditional on its realization, $\{a(k), \tau_y(k), \tau_\ell(k)\}$, the firm decides its production plan. If the entrant cannot profitably operate in any market, it will exit and forfeit the entry fee. Appendix B provides more details on the extended model with fixed costs. In the quantification part, the fixed costs of trade matrix f_{ij} are directly taken from Ma and Tang (2020), who measured these costs using the fraction of entrepreneurs in each prefecture in the *2005 One-Percent Population*

Table 4: Robustness Checks

(a) Semi-Elasticities

	Aggregate Welfare		Firm Entry		Net Flow to Rich Prefectures	
	σ	ρ	σ	ρ	σ	ρ
Baseline	-3.19	-0.84	-2.47	-0.49	3.91	6.01
$\psi = -0.2$	-3.21	-0.88	-2.47	-0.50	4.04	5.17
$\kappa = 1.5$	-3.22	-0.84	-2.49	-0.51	3.05	3.98
$\varepsilon = 5.0$	-2.30	-1.52	-1.32	-1.07	2.12	1.21
$f_{ii} > 0$	-3.21	-0.85	-2.49	-0.51	4.07	6.12
$f_{ij} > 0, \forall i, j$	-3.50	-0.79	-3.00	-0.31	4.95	5.78

(b) Effects of Migration

	Aggregate Welfare		Firm Entry		Gini Coefficient	
	σ	ρ	σ	ρ	σ	ρ
Baseline	5.83	9.16	6.59	24.18	-19.28	-32.65
$\psi = -0.2$	5.23	7.17	6.41	23.77	-18.74	-33.13
$\kappa = 1.5$	4.67	8.65	5.41	19.13	-15.98	-33.55
$\varepsilon = 5.0$	7.87	4.42	9.22	3.46	-16.10	-28.99
$f_{ii} > 0$	6.17	9.52	7.16	25.20	-18.37	-31.35
$f_{ij} > 0, \forall i, j$	5.61	-2.49	6.56	30.49	-21.39	-2.70

Notes: this table reports the robustness checks with respect to ψ, κ, ε, and f_{ij}. The numbers in the upper panel are the semi-elasticities of variables in the column headers to a reduction of 0.01 in both $\sigma_{y,j}$ and $\sigma_{\ell,j}$ under columns with header "σ". Similarly, the columns with the header "ρ" compute the semi-elasticity with a reduction of 0.01 in the absolute values of both $\rho_{ay,j}$ and $\rho_{a\ell,j}$. The lower panel reports the effects of migration, computed in the same way as in Table 3. We compute the Gini coefficient for the real wage. "Aggregate welfare" and "net flow to rich prefectures" are defined in the main text. The "number of entering firms" refers to "I" in the model. ψ is the dispersion elasticity; κ is the migration elasticity, and ε is the elasticity of substitution. In the table, setting $f_{ii} > 0$ allows selection into production but assumes away selection into exporting. $f_{ij} > 0$ allows selection into exporting.

Survey.

We present two sets of results to study the impacts of fixed costs. In the first, we only allow for $f_{ii} > 0$ and assume $f_{ij} = 0, \forall i \neq j$. This assumption introduces fixed operation costs so that firms with a total operating profit lower than f_{ii} exit the market. The first setup rules out selection into exporting so all the operating firms sell to all the markets. In the second, we assume $f_{ij} > 0$ for all i, j, enabling both selections into production and exporting.

Table 4 shows that the quantitative results are consistent between these versions and our baseline results. On the impacts of migration, market selection leads to mixed results. On the one hand, migration tends to widen the disparity in the number of varieties among locations. On the other hand, when migration is allowed the variations in factor prices across locations could be lower in the general equilibrium. Nevertheless, we observe similar results with the alternative specifications of fixed costs.

Correlation Parameter Across all the quantitative results, the effects of ρ are typically more sensitive than those of σ. As discussed in detail in the previous part, the reason is that positive and negative productivity correlations might have opposite welfare implications. As a result, reducing the absolute values of ρ could lead to counter-acting effects depending on the baseline estimates of ρ, rendering the results sensitive to other parameters.

6 Conclusion

This paper structurally estimates the within-prefecture distribution of firm-level frictions and studies their aggregate and distributional impacts. We show that the frictions in both factor and output markets vary systematically across prefectures. Firm-level frictions are less dispersed and correlated with productivity in richer prefectures. Our counterfactual exercise shows that reducing the within-prefecture misallocation increases aggregate welfare, discourages migration towards large prefectures, and narrows spatial income inequality. Moreover, we show that internal migration alleviates the aggregate impacts of micro-frictions and worsens spatial inequality simultaneously. Workers prefer to migrate out of the poor and heavily distorted locations to favor the richer and less distorted ones.

A couple of caveats exist in interpreting our results. Frictions are completely exogenous, and their causes must be rooted in certain institutional or geographical factors, which we do not explore in the current project. Our analysis focuses on manufacturing firms in the urban area; thus, our work cannot be directly applied to the agriculture and rural sectors. We also abstract away from the housing markets in the urban sectors and model the congestion through a stylized functional form. We focus on the basic patterns of within-prefecture misallocation, highlight their importance, and relay the abovementioned factors to future work.

References

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf, "The Economics of Density: Evidence from the Berlin Wall," *Econometrica*, 2015, *83* (6), 2127–2189.

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

Anderson, James E. and Eric van Wincoop, "Trade Costs," *Journal of Economic Literature*, September 2004, *42* (3), 691–751.

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

Bai, Yan, Keyu Jin, and Dan Lu, "Misallocation Under Trade Liberalization," Working Paper 26188, National Bureau of Economic Research August 2019.

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

_ , Gueorgui Kambourov, and Kjetil Storesletten, "Barriers to Entry and Regional Economic Growth in China," in "Conference on China's Financial Markets and the Global Economy, Suomen Pankki," Vol. 16 2017.

_ , Trevor Tombe, and Xiaodong Zhu, "Factor Market Distortions Across Time, Space and Sectors in China," *Review of Economic Dynamics*, 2013, *16* (1), 39–58.

Buera, Francisco J., Joseph P. Kaboski, and Yongseok Shin, "Finance and Development: A Tale of Two Sectors," *American Economic Review*, August 2011, *101* (5), 1964–2002.

Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro, "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock," *Econometrica*, 2019, pp. 741–835.

Fujita, Masahisa, Paul R Krugman, and Anthony Venables, *The spatial economy: Cities, regions, and international trade*, MIT press, 1999.

Gabaix, Xavier, "Zipf's Law and the Growth of Cities," *American Economic Review*, May 1999, *89* (2), 129–132.

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.

Guner, Nezih, Gustavo Ventura, and Yi Xu, "Macroeconomic Implications of Size-Dependent Policies," *Review of Economic Dynamics*, 2008, *11* (4), 721–744.

Hopenhayn, Hugo A., "On the Measure of Distortions," NBER Working Papers 20404, National Bureau of Economic Research, Inc August 2014.

Hsieh, Chang-Tai and Enrico Moretti, "Housing Constraints and Spatial Misallocation," *American Economic Journal: Macroeconomics*, April 2019, *11* (2), 1–39.

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

_ and Zheng Michael Song, "Grasp the Large, Let Go of the Small: The Transformation of the State Sector in China," *Brookings Papers on Economic Activity*, 2015, pp. 295–346.

Krugman, Paul, "Increasing Returns and Economic Geography," *Journal of Political Economy*, 1991, *99* (3), 483–499.

Ma, Lin and Yang Tang, "Geography, trade, and internal migration in China," *Journal of Urban Economics*, 2020, *115*, 103181. Cities in China.

_ and _ , "The Distributional Impacts of Transportation Networks in China," Technical Report, Available at SSRN 4118287 2022.

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

Midrigan, Virgiliu and Daniel Yi Xu, "Finance and Misallocation: Evidence from Plant-Level Data," *American Economic Review*, 2014, *104* (2), 422–458.

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

Restuccia, Diego and Richard Rogerson, "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments," *Review of Economic Dynamics*, 2008, *11* (4), 707–720.

Song, Zheng Michael and Guiying Laura Wu, "Identifying Capital Misallocation," Technical Report, working paper 2015.

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

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

Wu, Guiying Laura, "Capital misallocation in China: Financial frictions or policy distortions?," *Journal of Development Economics*, 2018, *130*, 203 – 223.

Yang, Mu-Jeung, "Micro-level Misallocation and Selection," *American Economic Journal: Macroeconomics*, October 2021, *13* (4), 341–68.

Appendix

A Additional Figures and Tables

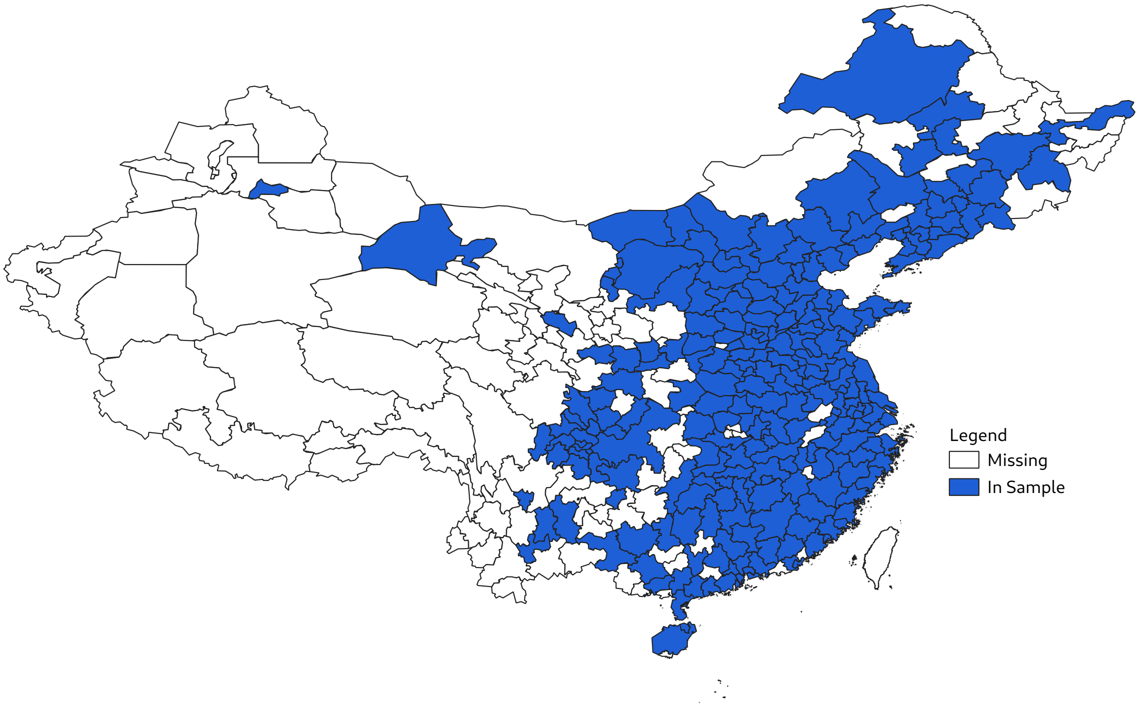
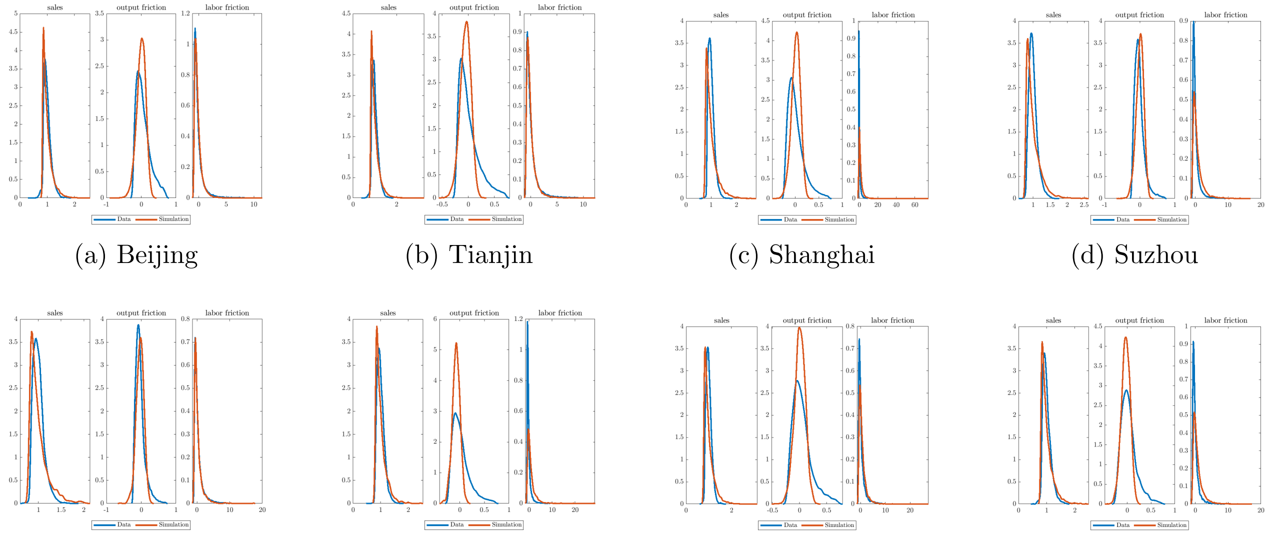


Figure A.1: Prefectures in the Sample

Notes: this map shows the 237 prefectures in the sample. We select the largest common set of prefectures between Ma and Tang (2020) and those with at least 500 unique firms in the ASIF database.



(a) Beijing

(b) Tianjin

(c) Shanghai

(d) Suzhou

(e) Guangzhou

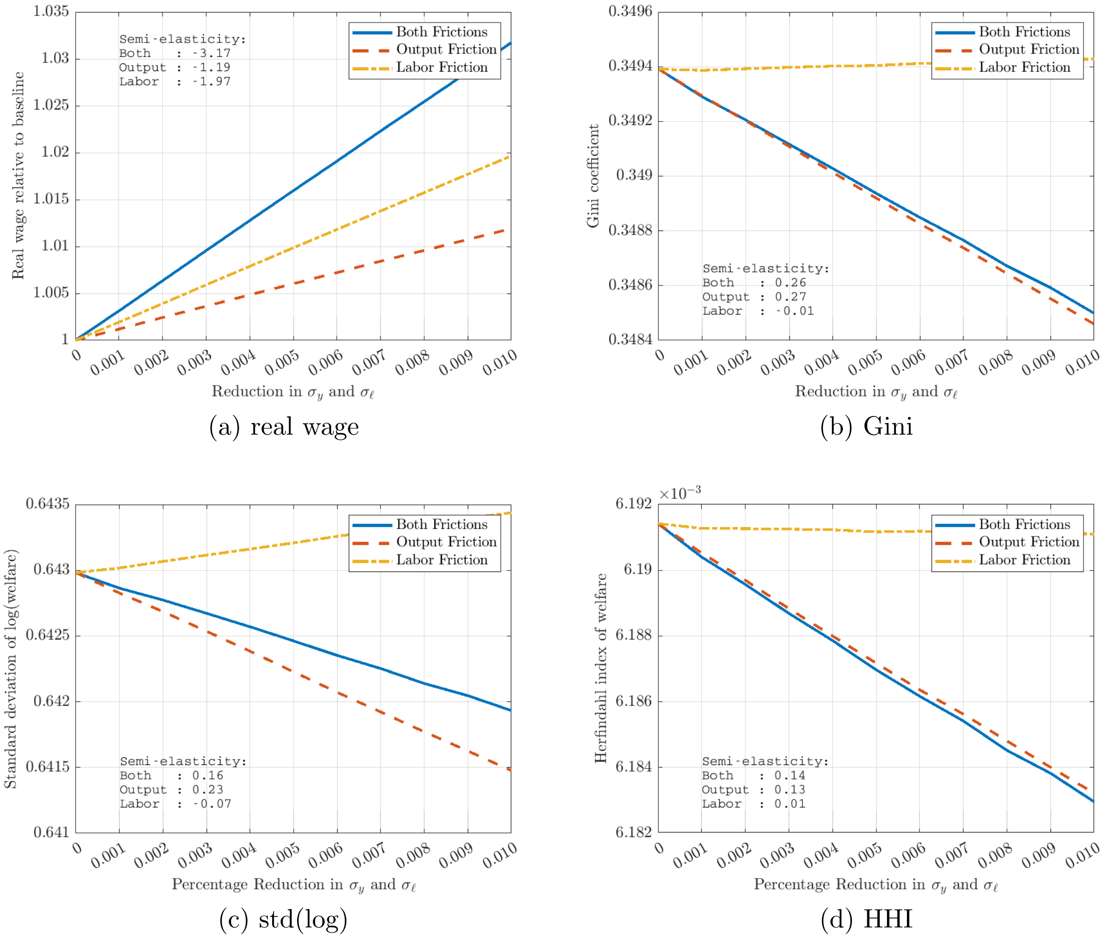
(f) Shenzhen

(g) Chengdu

(h) Xi’an

Figure A.2: Model Fit, Estimation

Notes: the eight panels present the kernel density estimates for the distribution of sales, output friction, and labor friction in both the data and the simulation for eight selected prefectures.



(a) real wage

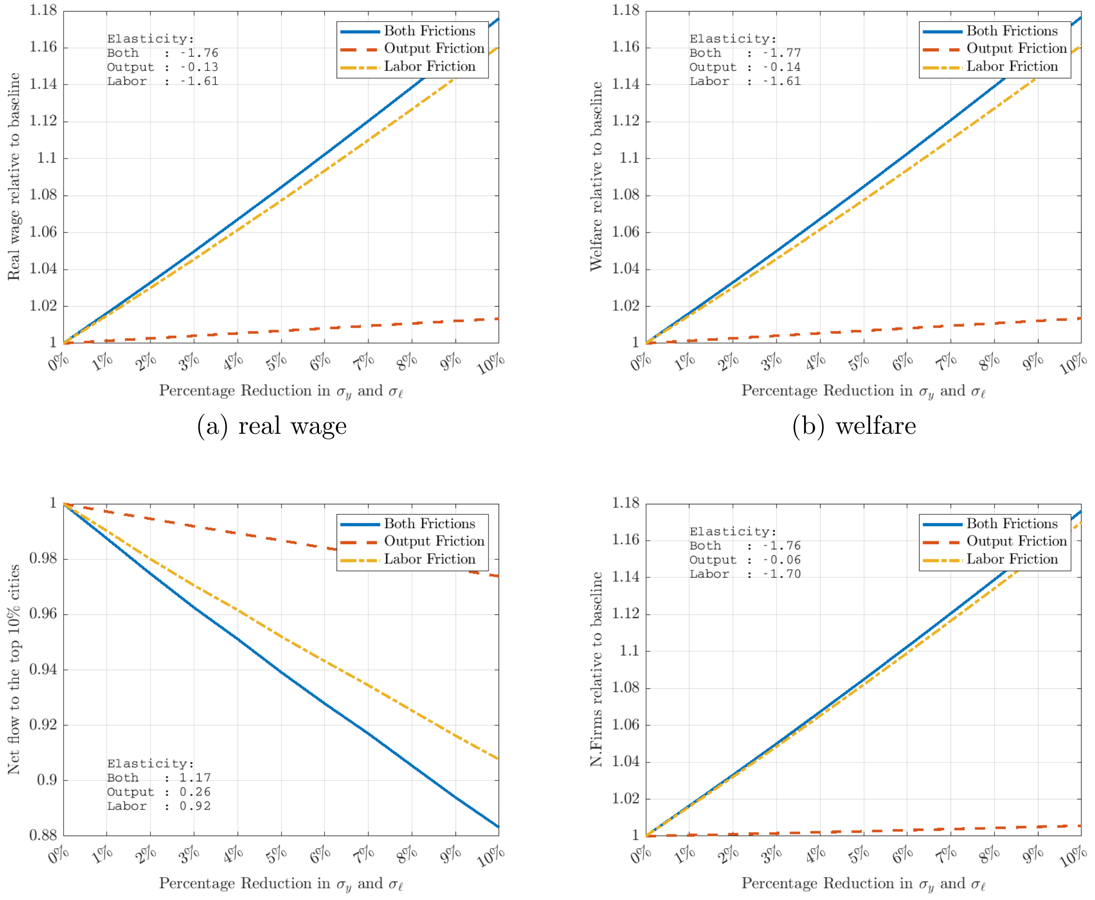
(b) Gini

(c) std(log)

(d) HHI

Figure A.3: Semi-Elasticity of $\sigma_{y,j}$ and $\sigma_{\ell,j}$, Additional Results

Notes: the four panels present the aggregate and distributional impacts of reducing the standard deviation of frictions in all the prefectures. The x-axes in all panels indicate the reduction in $\sigma_{y,j}$, and $\sigma_{\ell,j}$ in levels. The y-axes in panels (a) and (c) are the ratio of the real wage and the net population flows to the top 10% richest prefectures to the baseline level, respectively. In panel (b), the y-axis is the standard deviation of the natural logarithm of the welfare. In panel (d), the y-axis is the HHI index of welfare across prefectures.



(a) real wage

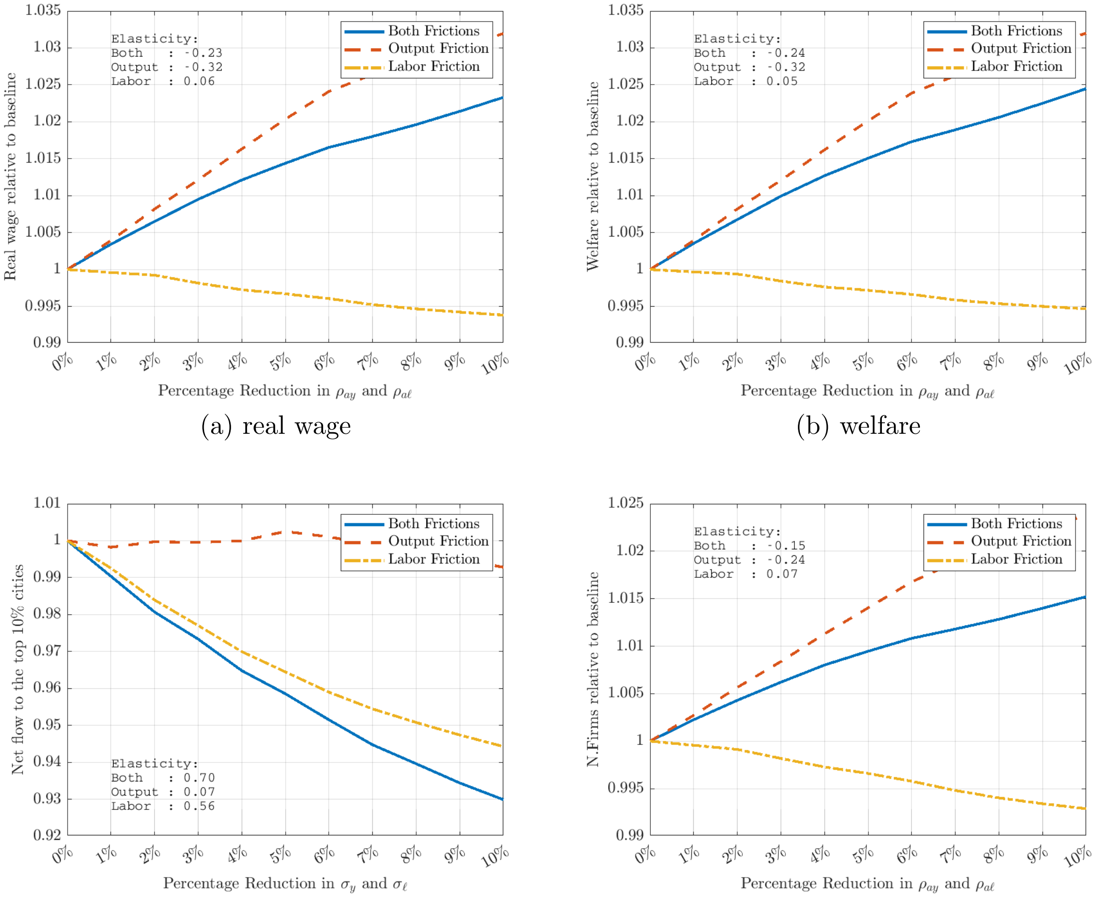
(b) welfare

(c) net flow to top prefectures by real wage

(d) firm entry

Figure A.4: Elasticity of $\sigma_{y,j}$ and $\sigma_{\ell,j}$

Notes: the four panels present the aggregate and distributional impacts of reducing the standard deviation of frictions in all the prefectures. The x-axes in all panels indicate the reduction in $\sigma_{y,j}$, and $\sigma_{\ell,j}$ in percentage terms.



(a) real wage

(b) welfare

(c) net flow to top prefectures by real wage

(d) firm entry

Figure A.5: Elasticity of $\rho_{ay,j}$ and $\rho_{a\ell,j}$

Notes: the four panels present the aggregate and distributional impacts of reducing the productivity correlations of frictions in all the prefectures. The x-axes indicate the reduction in $\rho_{ay,j}$, and $\rho_{a\ell,j}$ in percentage terms.

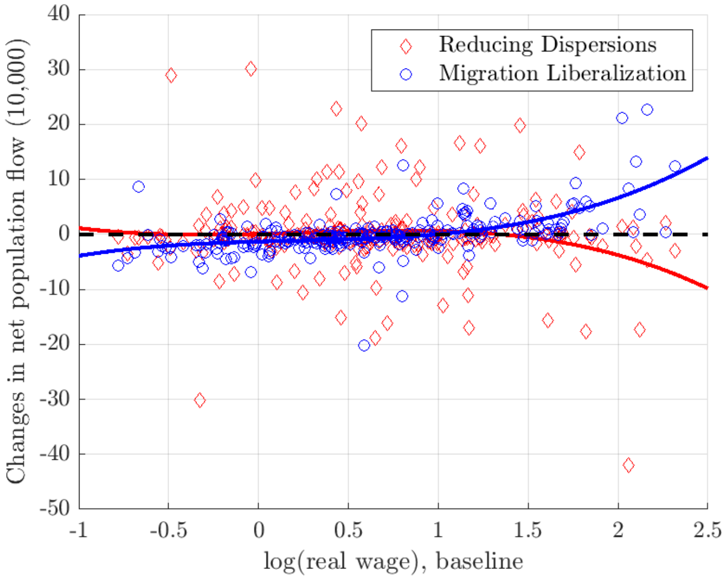


Figure A.6: Reducing correlation parameter v.s. migration frictions, Changes in Net Population Flow

Notes: the red and blue dots (curve) present the distributional impacts of reducing the productivity correlation parameters and migration liberalization in all the prefectures, respectively. The x-axis indicates the natural logarithm of the initial real wage. The y-axis is the log differences in the net population flow to receiving prefectures from baseline to the counterfactual economy.

Table A.1: Definition of Regions

Region	Provinces				
Southern Coast	Guangdong	Hainan			
Eastern Coast	Shanghai	Fujian	Jiangsu	Liaoning	
North	Beijing	Tianjin	Hebei	Heilongjiang	Jilin
	Liaoning	Shandong			
Others	-				

Notes: this table defines the "regions" used in the empirical analysis, sorted by average per-capita GDP. "Southern Coast" is the region with the highest per-capita GDP, followed by "Eastern Coast", "North", and "Others". All the provinces not included in the top three regions are in the "others" category. We classify all the prefectures in a province to the corresponding region.

Table A.2: Spatial Variation of Frictions across Regions

	(1)	(2)	(3)	(4)
	$\tilde{\sigma}_y$	$\tilde{\sigma}_\ell$	$\tilde{\rho}_{a,y}$	$\tilde{\rho}_{a,\ell}$
Southern Coast	-0.022***	-0.043	-0.289***	-0.015
	(0.002)	(0.042)	(0.032)	(0.025)
Eastern Coast	-0.020***	-0.160***	-0.299***	-0.028
	(0.003)	(0.031)	(0.039)	(0.023)
North	-0.002	0.108***	-0.043*	0.047***
	(0.003)	(0.030)	(0.022)	(0.016)
N	237	237	237	237
Adj.R-squared	0.217	0.160	0.391	0.044

Notes: this table regresses the dispersion and productivity correlation of output frictions ($\tilde{\sigma}_y$ and $\tilde{\rho}_{ay}$) and the labor frictions ($\tilde{\sigma}_\ell$ and $\tilde{\rho}_{a\ell}$) against regional dummy variables. The reference group is the "Others" region. See Table A.1 for the definition of regions. Robust standard errors are reported in the parenthesis. ***: significant at the 1% level; *: significant at the 5% level; *: significant at the 10% level.

B Solving the Model

In this section, we solve the extended model with the fixed costs. In particular, we assume that to "export" from location j to i, a fixed cost of f_{ij} in the unit of input bundle in location j is incurred. Upon entry, the firm draws its productivity and frictions from the joint distribution $G_j(a, \tau_y, \tau_\ell)$. Conditional on its realization, $\{a(k), \tau_y(k), \tau_\ell(k)\}$, the firm decides its production plan. If the entrant cannot profitably operate in any market, it will exit and forfeit the entry fee. The baseline model in the main text is a special case of the extended model with $f_{ij} = 0, \forall i, j$.

We need to solve a series of $\{w_j, I_j, P_j, X_j\}$ for all $j = 1, 2, \cdots, J$. Before providing the details of the solution, we first present the following results and notations that will be used extensively in the appendix. Note that the solution is based on the discretized version of the model, in which the frictions τ_y and τ_ℓ are discretized into grid points.

- We denote the firms in a grid point (τ_y^k, τ_ℓ^l) as "type d" firms, in which $d = 1, 2, \cdots, N_y \times N_\ell$ indexes the types.

- The CDF and the PDF of a in prefecture j, type d is:

$$G_j^d(a) = 1 - \left(\xi_j^d\right)^{\theta_j^d} a^{-\theta_j^d},$$
$$g_j^d(a) = \theta_j^d \left(\xi_j^d\right)^{\theta_j^d} a^{-\theta_j^d - 1},$$

  where ξ_j^d is the lower bound, and θ_j^d is the shape parameter. In practice, we found that $\xi_j^d = \bar{A}_j$, and thus directly enforce the condition.

- With $f_{ij} > 0$, the cut-off productivity is:

$$a_{ij}^d = \frac{\varepsilon}{\varepsilon - 1} \frac{t_{ij} c_j^d}{\left(1 - \tau_{y,j}^d\right) P_i} \left[\frac{\left(1 - \tau_{y,j}^d\right) X_i}{\varepsilon c_j^d f_{ij}}\right]^{\frac{1}{1-\varepsilon}}.$$

  In the baseline model with $f_{ij} = 0$, it is straightforward to see that $a_{ij}^d = \underline{a}$ for all i, j, and d.

- A potential entrant has probability λ_j^d to be type-d and the type is only revealed after paying the entry fee. Therefore:

$$I_j^d = \lambda_j^d I_j,$$
$$I_j = \sum_{d=1}^{D} I_j^d.$$

- The un-distorted cost of a bundle in prefecture j is:

$$\bar{c}_j = (1-\beta_j)^{\beta_j - 1} \beta_j^{-\beta_j} w_j^{\beta_j} P_j^{1-\beta_j},$$

and the distorted cost of a bundle is:

$$c_j^d = (1-\beta_j)^{\beta_j - 1} \beta_j^{-\beta_j} [(1+\tau_{\ell,j}^d) w_j]^{\beta_j} P_j^{1-\beta_j} = \left(1+\tau_{\ell,j}^d\right)^{\beta_j} \bar{c}_j.$$

- We use υ_{ij}^d to denote the following combination of parameters related to selling from j to i:

$$\upsilon_{ij}^d = \left(\frac{\varepsilon}{\varepsilon - 1}\right)^{1-\varepsilon} \left[\frac{t_{ij}}{\left(1-\tau_{y,j}^d\right)}\right]^{1-\varepsilon}. \tag{B.1}$$

and η_{ij}^d to denote the measure of trade flow from j to i:

$$\begin{aligned} \eta_{ij}^d &= \upsilon_{ij}^d \left(c_j^d\right)^{1-\varepsilon} \int_{a_{ij}^d}^{\infty} a^{\varepsilon-1} dG(a) \qquad \text{(B.2)} \\ &= \upsilon_{ij}^d \left(c_j^d\right)^{1-\varepsilon} \left(\frac{\left(\xi_j^d\right)^{\theta_j^d} \theta_j^d}{\theta_j^d - \varepsilon + 1}\right) \left(a_{ij}^d\right)^{-\theta_j^d + (\varepsilon - 1)} \end{aligned}$$

B.1 Updating Price

The ideal price index in prefecture i can thus be expressed as:

$$
\begin{aligned}
P_i^{1-\varepsilon} &= \sum_{j=1}^{J}\sum_{d=1}^{D} I_j^d \int_{a_{ij}^d}^{\infty} \left(\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c_j^d}{1-\tau_{y,j}^d}\right)^{1-\varepsilon} a^{\varepsilon-1} dG_j^d(a) \\
&= \sum_{j=1}^{J}\sum_{d=1}^{D} I_j^d \theta_j^d \left(\xi_j^d\right)^{\theta_j^d} \left(\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c_j^d}{1-\tau_{y,j}^d}\right)^{1-\varepsilon} \int_{a_{ij}^d}^{\infty} a^{\varepsilon-\theta_j^d-2} da \\
&= \sum_{j=1}^{J}\sum_{d=1}^{D} I_j^d \left(\xi_j^d\right)^{\theta_j^d} \left(\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c_j^d}{1-\tau_{y,j}^d}\right)^{1-\varepsilon} \left(\frac{\theta_j^d}{\theta_j^d-\varepsilon+1}\right) \left(a_{ij}^d\right)^{-\theta_j^d+(\varepsilon-1)} \\
P_i &= \left(\frac{\varepsilon}{\varepsilon-1}\right) \left\{ \sum_{j=1}^{J}\sum_{d=1}^{D} I_j^d \left(\frac{\left(\xi_j^d\right)^{\theta_j^d}\theta_j^d}{\theta_j^d-\varepsilon+1}\right) \left(\frac{t_{ij}}{1-\tau_{y,j}^d}\right)^{1-\varepsilon} \left(c_j^d\right)^{1-\varepsilon} \left(a_{ij}^d\right)^{-\theta_j^d+(\varepsilon-1)} \right\}^{\frac{1}{1-\varepsilon}}
\end{aligned}
$$

Alternatively, we can also express the price index using the definition of υ_{ij}^d:

$$
\begin{aligned}
P_i &= \left\{ \sum_{j=1}^{J}\sum_{d=1}^{D} \lambda_j^d I_j \upsilon_{ij}^d \left(c_j^d\right)^{1-\varepsilon} \int_{a_{ij}^d}^{\infty} a^{\varepsilon-1} dG(a) \right\}^{\frac{1}{1-\varepsilon}} \\
&= \left\{ \sum_{j=1}^{J}\sum_{d=1}^{D} \lambda_j^d I_j \eta_{ij}^d \right\}^{\frac{1}{1-\varepsilon}} \qquad \text{(B.3)}
\end{aligned}
$$

B.2 Updating Wage

We back out the wage rates using labor market clearing conditions. The pre-tax sales revenue from i to j by type-d is:

$$
\begin{aligned}
X_{ji}^d &= I_i^d \int_{a_{ji}^d}^{\infty} \frac{X_j}{P_j^{1-\varepsilon}} \left[p_{ji}^d(a)\right]^{1-\varepsilon} dG(a) \\
&= I_i^d \int_{a_{ji}^d}^{\infty} \frac{X_j}{P_j^{1-\varepsilon}} \left(\frac{\varepsilon}{\varepsilon-1}\frac{t_{ji}c_i^d}{1-\tau_{y,i}^d}\right)^{1-\varepsilon} a^{\varepsilon-1} dG(a) \\
&= I_i^d \frac{X_j}{P_j^{1-\varepsilon}} \left(\frac{\varepsilon}{\varepsilon-1}\frac{t_{ji}c_i^d}{1-\tau_{y,i}^d}\right)^{1-\varepsilon} \frac{\left(\xi_j^d\right)^{\theta_j^d}\theta_j^d}{\theta_j^d-(\varepsilon-1)} \left(a_{ji}^d\right)^{-\theta_j^d+(\varepsilon-1)}. \qquad \text{(B.4)}
\end{aligned}
$$

Alternatively, the trade flow can also be expressed as:

$$X^d_{ji} = I^d_i \frac{X_j}{P_j^{1-\varepsilon}} \eta^d_{ji}.$$

Note that:

$$X^d_i = \sum_{j=1}^{J} X^d_{ji},$$

is the total revenue of the type-d firms in prefecture i.

The number of input bundles required to generate the sales above is:

$$\begin{aligned}
B^d_{ji} &= I^d_i \int_{a^d_{ji}}^{\infty} t_{ji} \frac{q^d_{ji}}{a} dG(a) \\
&= I^d_i \int_{a^d_{ji}}^{\infty} \frac{X_j}{P_j^{1-\varepsilon}} \left(\frac{\varepsilon}{\varepsilon - 1} \frac{t_{ji} c^d_i}{1 - \tau^d_{y,i}} \frac{1}{a} \right)^{-\varepsilon} \frac{t_{ji}}{1 - \tau^d_{y,i}} \left(1 - \tau^d_{y,i}\right) \frac{1}{a} dG(a) \\
&= I^d_i \frac{\left(1 - \tau^d_{y,i}\right) X_j}{P_j^{1-\varepsilon}} \left(\frac{t_{ji}}{1 - \tau^d_{y,i}} \right)^{1-\varepsilon} \left(\frac{\varepsilon}{\varepsilon - 1} c^d_i \right)^{-\varepsilon} \int_{a^d_{ji}}^{\infty} a^{\varepsilon - 1} dG(a) \\
&= I^d_i \frac{\left(1 - \tau^d_{y,i}\right) X_j}{P_j^{1-\varepsilon}} \left(\frac{t_{ji}}{1 - \tau^d_{y,i}} \right)^{1-\varepsilon} \left(\frac{\varepsilon}{\varepsilon - 1} c^d_i \right)^{-\varepsilon} \frac{\left(\xi^d_i\right)^{\theta^d_i} \theta^d_i}{\theta^d_i - (\varepsilon - 1)} \left(a^d_{ji}\right)^{-\theta^d_i + (\varepsilon - 1)}.
\end{aligned}$$

Alternatively, the bundle requirement can be expressed as:

$$\begin{aligned}
B^d_{ji} &= I^d_i \frac{\left(1 - \tau^d_{y,i}\right) X_j}{P_j^{1-\varepsilon}} \frac{\left(\frac{t_{ji}}{1 - \tau^d_{y,i}} \right)^{1-\varepsilon} \left(\frac{\varepsilon}{\varepsilon - 1} c^d_i \right)^{1-\varepsilon}}{\left(\frac{\varepsilon}{\varepsilon - 1} c^d_i \right)} \frac{\left(\xi^d_i\right)^{\theta^d_i} \theta^d_i}{\theta^d_i - (\varepsilon - 1)} \left(a^d_{ji}\right)^{-\theta^d_i + (\varepsilon - 1)} \\
&= I^d_i \frac{\left(1 - \tau^d_{y,i}\right) X_j}{P_j^{1-\varepsilon}} \frac{\nu^d_{ji} \left(c^d_i\right)^{1-\varepsilon}}{\left(\frac{\varepsilon}{\varepsilon - 1} c^d_i \right)} \frac{\left(\xi^d_i\right)^{\theta^d_i} \theta^d_i}{\theta^d_i - (\varepsilon - 1)} \left(a^d_{ji}\right)^{-\theta^d_i + (\varepsilon - 1)}. \\
&= I^d_i \frac{\left(1 - \tau^d_{y,i}\right) X_j}{P_j^{1-\varepsilon}} \frac{\eta^d_{ji}}{\left(\frac{\varepsilon}{\varepsilon - 1} c^d_i \right)}
\end{aligned}$$

The relationship between B^d_{ji} and X^d_{ji} is therefore:

$$X^d_{ji} = B^d_{ji} c^d_i \frac{\varepsilon}{\varepsilon - 1} \frac{1}{1 - \tau^d_{y,i}}.$$

For each unit of input bundle, the labor demand that minimizes the unit costs is:

$$\begin{aligned} l_i^d &= \left[\frac{P_i}{(1+\tau_{\ell,i}^d)w_i}\right]^{1-\beta_j} \left(\frac{\beta_j}{1-\beta_j}\right)^{1-\beta_j} \\ &= \frac{\beta_j c_i^d}{(1+\tau_{\ell,i}^d)w_i}, \end{aligned}$$

and thus the labor demand to generate X_{ji} is $l_i^d \cdot B_{ji}^d$. Firms also need to purchase input bundles to pay for the fixed operating costs. The total demand for input bundles in prefecture i in order to cover the operating costs of serving prefecture j is:

$$\begin{aligned} B_{f,ji}^d &= f_{ji} I_i^d \int_{a_{ji}^d}^{\infty} dG(a) = f_{ji} I_i^d \cdot \left(1 - G(a_{ji}^d)\right) \\ &= f_{ji} I_i^d \left(\xi_i^d\right)^{\theta_i^d} \left(a_{ji}^d\right)^{-\theta_i^d}. \end{aligned}$$

As a result, the total labor demand incurred in prefecture i for serving prefecture j is the sum of both variable and fixed costs of production:

$$L_{ji}^d = l_i^d \cdot (B_{ji}^d + B_{f,ji}^d).$$

The total labor demand in prefecture i is thus the sum of labor demand by all destinations and types of firms and the labor incurred to cover the entry costs measured as undistorted input bundles. Finally, the labor market clearing condition in prefecture i is given as:

$$L_i = \sum_{j=1}^{N} \sum_{d=1}^{D} L_{ji}^d + I_i f_e \left(\frac{P_i}{w_i}\right)^{1-\beta_j} \left(\frac{\beta_j}{1-\beta_j}\right)^{1-\beta_j},$$

where the LHS is the total labor supply in prefecture i.

B.3 Updating Expenditure

The total expenditure of prefecture j, X_j, is the final spending by the consumers, plus expenditure on composite varieties:

$$X_j = Y_j + (1-\beta_j)\left[\sum_{d=1}^{D} c_j^d \sum_{i=1}^{J}\left(B_{ij}^d + B_{f,ij}^d\right) + I_j f_e \bar{c}_j\right]. \tag{B.5}$$

Different from the labor-market clearing condition, $(1-\beta_j)$ fraction of the total bundle costs regardless of distortion, $\left[\sum_{d=1}^{D} c_j^d \sum_{i=1}^{J}\left(B_{ij}^d + B_{f,ij}^d\right) + I_j f_e \bar{c}_j\right]$, must be the expenditure on the differentiated products, as there is no additional wedge in intermediate goods usage.

The equilibrium income in prefecture j, denoted as Y_j, includes both the labor income and the lump-sum transfer/tax from both distortions:

$$Y_j = w_j L_j + \sum_{d=1}^{D} \tau_{y,j}^d X_j^d + \sum_{d=1}^{D} \frac{\beta_j \tau_{\ell,j}^d}{1+\tau_{\ell,j}^d} c_j^d \sum_{i=1}^{J}\left(B_{ij}^d + B_{f,ij}^d\right). \tag{B.6}$$

In the expression above, the first term is the total labor income, the second is the total tax revenue from the output wedge, and the last is the labor wedge. Similar to the expression in the previous section, $\beta_j c_j^d \sum_{i=1}^{J}\left(B_{ij}^d + B_{f,ij}^d\right)$ is the total payroll costs, net of the undistorted entry, to the firms. Out of distorted labor costs, $\frac{\tau_{\ell,j}^d}{1+\tau_{\ell,j}^d}$ fraction is the payroll tax. Substitute the expression of Y_j into the expression of X_j:

$$\begin{aligned} X_j = w_j L_j &+ \sum_{d=1}^{D} \tau_{y,j}^d X_j^d + \sum_{d=1}^{D} \frac{\beta_j \tau_{\ell,j}^d}{1+\tau_{\ell,j}^d} c_j^d \sum_{i=1}^{J}\left(B_{ij}^d + B_{f,ij}^d\right) \\ &+ (1-\beta_j)\left[\sum_{d=1}^{D} c_j^d \sum_{i=1}^{J}\left(B_{ij}^d + B_{f,ij}^d\right) + I_j f_e \bar{c}_j\right]. \end{aligned}$$

B.4 Updating the Number of Entrants

The free entry condition in prefecture j is:

$$\sum_{i=1}^{J}\sum_{d=1}^{D}\left\{\lambda_j^d \int_{a_{ij}^d}^{\infty} \frac{(1-\tau_{y,j}^d)X_i}{\varepsilon P_i^{1-\varepsilon}}\left(\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c_j^d}{1-\tau_{y,j}^d}\right)^{1-\varepsilon} a^{\varepsilon-1} - c_j^d f_{ij} dG(a)\right\} = f_e \bar{c}_j. \tag{B.7}$$

In the equation above, the left-hand side is the expected profit before the realization of type and productivity, and the right-hand side is the cost of entry. Re-arrange the equation above using the definition of υ_{ij} in the equation (B.1):

$$\sum_{i=1}^{J}\frac{X_i}{\varepsilon P_i^{1-\varepsilon}}\sum_{d=1}^{D}\lambda_j^d\left(1-\tau_{y,j}^d\right)\upsilon_{ij}^d\left(c_j^d\right)^{1-\varepsilon}\int_{a_{ij}^d}^{\infty}a^{\varepsilon-1}dG(a)=f_e\bar{c}_j+\sum_{i=1}^{J}\sum_{d=1}^{D}\lambda_j^d\int_{a_{ij}^d}^{\infty}c_j^d f_{ij}dG(a).$$

Stack the above equation for all the origin prefectures $j = 1, \cdots, J$ into a matrix form:

$$\begin{bmatrix}\sum_{d=1}^{D}\lambda_1^d\left(1-\tau_{y,1}^d\right)\eta_{11}^d & \sum_{d=1}^{D}\lambda_1^d\left(1-\tau_{y,1}^d\right)\eta_{21}^d & \cdots & \sum_{d=1}^{D}\lambda_1^d\left(1-\tau_{y,1}^d\right)\eta_{J1}^d\\ \sum_{d=1}^{D}\lambda_2^d\left(1-\tau_{y,2}^d\right)\eta_{12}^d & \sum_{d=1}^{D}\lambda_2^d\left(1-\tau_{y,2}^d\right)\eta_{22}^d & \cdots & \sum_{d=1}^{D}\lambda_2^d\left(1-\tau_{y,2}^d\right)\eta_{J2}^d\\ \vdots & \vdots & \vdots & \vdots\\ \sum_{d=1}^{D}\lambda_J^d\left(1-\tau_{y,J}^d\right)\eta_{1J}^d & \sum_{d=1}^{D}\lambda_J^d\left(1-\tau_{y,J}^d\right)\eta_{2J}^d & \cdots & \sum_{d=1}^{D}\lambda_J^d\left(1-\tau_{y,J}^d\right)\eta_{JJ}^d\end{bmatrix}$$

$$\times\begin{bmatrix}\frac{X_1}{\varepsilon}P_1^{\varepsilon-1}\\ \frac{X_2}{\varepsilon}P_2^{\varepsilon-1}\\ \vdots\\ \frac{X_J}{\varepsilon}P_J^{\varepsilon-1}\end{bmatrix}=\begin{bmatrix}f_e\bar{c}_1+\sum_{i=1}^{J}\sum_{d=1}^{D}\lambda_1^d\int_{a_{i1}^d}^{\infty}c_1^d f_{i1}dG(a)\\ f_e\bar{c}_2+\sum_{i=1}^{J}\sum_{d=1}^{D}\lambda_2^d\int_{a_{i2}^d}^{\infty}c_2^d f_{i2}dG(a)\\ \vdots\\ f_e\bar{c}_J+\sum_{i=1}^{J}\sum_{d=1}^{D}\lambda_J^d\int_{a_{iJ}^d}^{\infty}c_J^d f_{iJ}dG(a)\end{bmatrix}.$$

Denote the matrix in the first line as Ψ, and substitute in the solution of price in equation (B.3):

$$\begin{bmatrix}\frac{X_1}{\varepsilon}\left\{\sum_{j=1}^{J}\sum_{d=1}^{D}\lambda_j^d I_j\upsilon_{1j}^d\left(c_j^d\right)^{1-\varepsilon}\int_{a_{1j}^d}^{\infty}a^{\varepsilon-1}dG(a)\right\}^{-1}\\ \frac{X_2}{\varepsilon}\left\{\sum_{j=1}^{J}\sum_{d=1}^{D}\lambda_j^d I_j\upsilon_{2j}^d\left(c_j^d\right)^{1-\varepsilon}\int_{a_{2j}^d}^{\infty}a^{\varepsilon-1}dG(a)\right\}^{-1}\\ \vdots\\ \frac{X_J}{\varepsilon}\left\{\sum_{j=1}^{J}\sum_{d=1}^{D}\lambda_j^d I_j\upsilon_{Jj}^d\left(c_j^d\right)^{1-\varepsilon}\int_{a_{Jj}^d}^{\infty}a^{\varepsilon-1}dG(a)\right\}^{-1}\end{bmatrix}=\left(\Psi\right)^{-1}\begin{bmatrix}f_e\bar{c}_1+\sum_{i=1}^{J}\sum_{d=1}^{D}\lambda_1^d\int_{a_{i1}^d}^{\infty}c_1^d f_{i1}dG(a)\\ f_e\bar{c}_2+\sum_{i=1}^{J}\sum_{d=1}^{D}\lambda_2^d\int_{a_{i2}^d}^{\infty}c_2^d f_{i2}dG(a)\\ \vdots\\ f_e\bar{c}_J+\sum_{i=1}^{J}\sum_{d=1}^{D}\lambda_J^d\int_{a_{iJ}^d}^{\infty}c_J^d f_{iJ}dG(a)\end{bmatrix}.$$

where $\left(\Psi\right)^{-1}$ is the inverse of the matrix Ψ. Denote the RHS vector of the above equation as $\vec{F}$, and with the understanding that $\vec{F}_j$ is the jth element of $\vec{F}$, we can re-arrange the equation as:

$$\begin{bmatrix} \frac{\varepsilon}{X_1}\sum_{j=1}^{J} I_j \sum\limits_{d=1}^{D} \lambda_j^d \eta_{1j}^d \\ \frac{\varepsilon}{X_2}\sum_{j=1}^{J} I_j \sum\limits_{d=1}^{D} \lambda_j^d \eta_{2j}^d \\ \vdots \\ \frac{\varepsilon}{X_3}\sum_{j=1}^{J} I_j \sum\limits_{d=1}^{D} \lambda_j^d \eta_{Jj}^d \end{bmatrix} = \begin{bmatrix} \frac{1}{\vec{F}_1} \\ \frac{1}{\vec{F}_2} \\ \vdots \\ \frac{1}{\vec{F}_J} \end{bmatrix}.$$

Note that in the equation above, we have used the definition of η_{ij}^d defined in equation (B.2). Re-write the LHS of the above equation as a matrix multiplication:

$$\begin{bmatrix} \sum_{d=1}^{D} \lambda_1^d \eta_{11}^d & \sum_{d=1}^{D} \lambda_2^d \eta_{12}^d & \cdots & \sum_{d=1}^{D} \lambda_J^d \eta_{1J}^d \\ \sum_{d=1}^{D} \lambda_1^d \eta_{21}^d & \sum_{d=1}^{D} \lambda_2^d \eta_{22}^d & \cdots & \sum_{d=1}^{D} \lambda_J^d \eta_{2J}^d \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{d=1}^{D} \lambda_1^d \eta_{J1}^d & \sum_{d=1}^{D} \lambda_2^d \eta_{J2}^d & \cdots & \sum_{d=1}^{D} \lambda_J^d \eta_{JJ}^d \end{bmatrix} \times \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_J \end{bmatrix} = \begin{bmatrix} \frac{X_1}{\varepsilon \vec{F}_1} \\ \frac{X_2}{\varepsilon \vec{F}_2} \\ \vdots \\ \frac{X_J}{\varepsilon \vec{F}_J} \end{bmatrix}.$$

Denote the LHS matrix on the first line as Φ, and the number of entrants is computed as:

$$\begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_J \end{bmatrix} = (\Phi)^{-1} \begin{bmatrix} \frac{X_1}{\varepsilon \vec{F}_1} \\ \frac{X_2}{\varepsilon \vec{F}_2} \\ \vdots \\ \frac{X_J}{\varepsilon \vec{F}_J} \end{bmatrix}. \tag{B.8}$$

The above solution based on matrix inversion is fast but unstable. If the linear solution fails, we then directly solve equation (B.7) as a system of non-linear equations using gradient-based methods.

C Quantification

C.1 Marginal Distributions

In each prefecture, $\log(1-\tau_y)$, $\log(1+\tau_\ell)$, and a follow a joint distribution, in which the marginal distributions of $\log(1-\tau_y)$ and $\log(1+\tau_\ell)$ are Gaussian with a normalized mean of zero, and the marginal distribution of a is Pareto. We also allow for a general correlation structure across the three dimensions, as captured by the three correlation parameters, $\rho_{y,\ell}$, $\rho_{y,a}$, and $\rho_{\ell,a}$. We need to estimate the following 6 parameters: the 3 correlations, the shape parameter governing the Pareto productivity, the 2 standard deviations of τ_y, and τ_ℓ.

We assume that we can observe the top x percent of the firms sorted by augmented productivity, as defined below.

Augmented productivity and the ranking of firms The revenue of firm k originating in prefecture j and selling to prefecture i with a draw of $a, \tau^d_{y,j}, \tau^d_{\ell,j}$ is:

$$\begin{aligned} r^d_{ij} &= \frac{\left(1-\tau^d_{y,j}\right) X_i}{P_i^{1-\varepsilon}} \left[\frac{\varepsilon}{\varepsilon-1}\frac{t_{ij}c^d_j}{\left(1-\tau^d_{y,j}\right)a}\right]^{1-\varepsilon} \\ &= \left(\frac{\varepsilon t_{ij}}{\varepsilon-1}\right)^{1-\varepsilon}\left[\frac{X_i}{P_i^{1-\varepsilon}}\right]\left\{a^{\varepsilon-1}\left(1-\tau^d_{y,j}\right)^{\varepsilon}\left[\left(1+\tau^d_{\ell,j}\right)^{\beta_j}\bar{c}_j\right]^{1-\varepsilon}\right\} \\ &= \left(\frac{\varepsilon t_{ij}\bar{c}_j}{\varepsilon-1}\right)^{1-\varepsilon}\left[\frac{X_i}{P_i^{1-\varepsilon}}\right]\left\{a^{\varepsilon-1}\left(1-\tau^d_{y,j}\right)^{\varepsilon}\left(1+\tau^d_{\ell,j}\right)^{\beta_j(1-\varepsilon)}\right\}. \end{aligned}$$

In the expression above, the first two terms are common across all the firms, and the firms ranked by the "augmented productivity" in the curly bracket, defined as $\tilde{a}^d_j$:

$$\left(\tilde{a}^d_j\right)^{\varepsilon-1} = a^{\varepsilon-1}\left(1-\tau^d_{y,j}\right)^{\varepsilon}\left(1+\tau^d_{\ell,j}\right)^{\beta_j(1-\varepsilon)}.$$

It is straightforward to see that a firm with higher $\tilde{a}^d_j$ will be larger and more productive. In a random draw of firms, the firms with the highest $\tilde{a}^d_j$ will survive regardless of t_{ij} and f_{ij}.

Aggregating the expression of sales across different markets, the revenue of the firm is:

$$\begin{aligned}\sum_{i=1}^{J} r_{ij}^d &= \sum_{i=1}^{J} \left(\frac{\varepsilon t_{ij}\bar{c}_j}{\varepsilon - 1}\right)^{1-\varepsilon} \left[\frac{X_i}{P_i^{1-\varepsilon}}\right] \left\{a^{\varepsilon-1}\left(1-\tau_{y,j}^d\right)^{\varepsilon}\left(1+\tau_{\ell,j}^d\right)^{\beta_j(1-\varepsilon)}\right\} \\ &= \left(\frac{\varepsilon \bar{c}_j}{\varepsilon - 1}\right)^{1-\varepsilon} \left(\tilde{a}_j^d\right)^{\varepsilon-1} \left\{\sum_{i=1}^{J} \left(t_{ij}\right)^{1-\varepsilon} \frac{X_i}{P_i^{1-\varepsilon}}\right\}.\end{aligned}$$

The terms in the curly bracket summarize the markets the firm sells. The market access term is firm-specific, as more productive firms will break into more markets.

C.2 Conditional Distributions

We discretize τ_y and τ_ℓ onto a grid, and on each grid point, we need to estimate the conditional distribution of a. We do this by Monte Carlo simulations as the conditional distribution from Gaussian copulas does not adopt closed-form solutions. We verify that the conditional distributions are similar to Pareto using Zipf plots.

Degenerate Distributions In several cases, the estimated conditional distribution leads to very high levels of θ_j^d. Numerically the resulting $(\xi_j^d)^{\theta_j^d} \approx \infty$. Theoretically, as $\theta_j^d \to \infty$, it is straightforward to see that the Pareto distribution collapses to a degenerate distribution at ξ_j^d. In light of this, we approximate the conditional distribution as a degenerate distribution to avoid numerical infinities in implementation.

In the case of a degenerate distribution, the following integration will be affected:

$$\int_{a_{ij}^d}^{\infty} a^{\varepsilon-1} dG(a) = \begin{cases} 0 & , \text{ if } a_{ij}^d > \xi_j^d \\ \left(\xi_j^d\right)^{\varepsilon-1} & , \text{ if } a_{ij}^d \leq \xi_j^d \end{cases}$$

This integration shows up in the expression of η_{ij}^d, P_i,X_{ij}^d, and B_{ij}^d. In addition, the probability of entry becomes binary (0 or 1), and therefore the following variables will be affected: $B_{f,ji}^d$ and I_j^d. In particular,

$$\int_{a_{ij}^d}^{\infty} dG(a) = 1 - G(a_{ij}^d) = \begin{cases} 0 & , \text{ if } a_{ij}^d > \xi_j^d \\ 1 & , \text{ if } a_{ij}^d \leq \xi_j^d \end{cases}$$

C.3 Joint Calibration

We solve the joint calibration problem as a fixed point problem and use the Gauss-Jacobi algorithm to find the fixed point. The algorithm ends with a tolerance value of `1.0E-4`.

D Additional Results

D.1 Regional Variations of Frictions

Table D.1 regresses the standard deviation and productivity correlation of output and labor frictions against per-capita GDP by gradually controlling prefecture-level characteristics, such as population, labor share, the share of the manufacturing industry in GDP, and connectivity in the transportation network. A negative and significant relation between the dispersion of output frictions and income persists in all our regressions. A one-percent increase in per capita GDP decreases the standard deviation of output frictions by 0.006. These results imply that richer prefectures tend to have less dispersed output frictions. In addition, locations with higher labor share in production costs or those more remote cities seem to experience higher dispersion of output frictions. We have also explored how correlations with productivity vary against income. The productivity correlations of output frictions also decrease with income: a one-percent increase in per capita GDP leads to a decrease of 0.107 in the correlations.

The relationship between income and the dispersion of labor frictions is also negative. Our estimation indicates that a one-percent increase in per capita GDP decreases $\tilde{\sigma}_\ell$. However, this coefficient is not precisely identified. Those prefectures with a higher labor share in production experience less dispersed labor frictions. The relationship between the productivity correlation of labor frictions and income is also significantly negative: a one-percent increase in per capita GDP reduces the productivity correlation by 0.028.

Table D.1: Spatial Distribution of Frictions

(a) Output Frictions

	$\tilde{\sigma}_y$					$\tilde{\rho}_{ay}$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
log(per-capita GDP)	-0.004**	-0.005***	-0.005**	-0.004*	-0.006**	-0.092***	-0.098***	-0.107***	-0.102***	-0.107***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.020)	(0.020)	(0.023)	(0.024)	(0.025)
log(population)		-0.007***	-0.007***	-0.006***	-0.003*		-0.049***	-0.045**	-0.042**	-0.034*
		(0.002)	(0.002)	(0.002)	(0.002)		(0.018)	(0.019)	(0.018)	(0.020)
industry share			0.000	-0.000	0.000			0.001	0.001	0.002
			(0.000)	(0.000)	(0.000)			(0.001)	(0.001)	(0.002)
labor share				0.228***	0.254***				1.022	1.095
				(0.084)	(0.080)				(0.827)	(0.816)
remoteness					0.090***					0.251
					(0.020)					(0.164)
N	237	237	237	237	237	237	237	237	237	237
Adj.R-squared	0.015	0.066	0.062	0.083	0.158	0.087	0.109	0.109	0.110	0.112

(b) Labor Frictions

	$\tilde{\sigma}_\ell$					$\tilde{\rho}_{a\ell}$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
log(per-capita GDP)	0.002	0.000	-0.003	-0.017	-0.024	-0.013	-0.012	-0.035***	-0.036***	-0.028**
	(0.022)	(0.022)	(0.023)	(0.023)	(0.024)	(0.010)	(0.011)	(0.013)	(0.014)	(0.013)
log(population)		-0.018	-0.017	-0.023	-0.012		0.012	0.024*	0.024*	0.011
		(0.022)	(0.022)	(0.023)	(0.025)		(0.012)	(0.012)	(0.013)	(0.014)
industry share			0.000	0.001	0.002			0.003***	0.003***	0.002***
			(0.001)	(0.001)	(0.001)			(0.001)	(0.001)	(0.001)
labor share				-2.963***	-2.864**				-0.054	-0.168
				(1.107)	(1.124)				(0.476)	(0.477)
remoteness					0.341					-0.391***
					(0.231)					(0.123)
N	237	237	237	237	237	237	237	237	237	237
Adj.R-squared	-0.004	-0.005	-0.009	0.019	0.023	0.001	0.002	0.066	0.062	0.098

Notes: this table reports the regression results of the dispersion and productivity correlation of output and labor frictions against prefecture characteristics. Robust standard errors are reported in the parenthesis. ***: significant at the 1% level; *: significant at the 5% level; *: significant at the 10% level. "Industry share" refers to the share of the manufacturing industry in GDP, "labor share" refers to β_j discussed earlier, and "remoteness" of a prefecture measures the location of a prefecture in the transportation network, from Ma and Tang (2020).

D.2 Non-Zero Mean in Friction Distributions

In the baseline model, we assume that the marginal distributions of frictions follow log-normal distributions with zero means. In this part, we relax this assumption and present the results with non-zero means. In particular, we assume that the marginal distributions of the frictions are:

$$\begin{aligned}\log(1-\tau_y) &\sim \mathcal{N}(\mu_{y,j}, \sigma_{y,j}) \\ \log(1+\tau_\ell) &\sim \mathcal{N}(\mu_{\ell,j}, \sigma_{\ell,j}),\end{aligned}$$

where $\{\mu_{y,j}, \mu_{\ell,j}\}$ are the mean of the log-normal distributions, which are new parameters that we need to estimate for each prefecture. The additional moments we try to match are the mean of $\log(1-\tau_y)$ and $\log(1+\tau_\ell)$ given our estimated $\{\tau_y, \tau_\ell\}$ from the data. We repeat the estimation and quantification procedure as described in the main text, with the difference that now, each prefecture is characterized by eight instead of six parameters.

The estimated dispersion and correlation parameters highly correlate with those in the baseline specification. Table D.2 regresses the estimated parameters under non-zero mean assumptions against those in the baseline estimation. As shown in the first four columns of the table, all coefficients are positive, although the relationship between the correlation parameters is imprecisely measured. The similarity of these parameters between the baseline and the extended model suggests that the main results in the baseline model are unlikely to be driven by the zero-mean assumption.

The estimated mean parameters correlate with the dispersion parameters, as shown in the last two columns of the same table. In other words, locations with a higher dispersion of firm-level frictions are also likely the locations with higher average frictions. The co-movement between the mean and the dispersion of frictions is expected, as the distribution of dispersions is likely influenced by location-specific institutional quality.

To understand the welfare implications of the estimated frictions, we repeat the exercises reported in the main text. Recall that to quantify the model, we first re-estimate the conditional productivity distributions and re-calibrate the parameters reported in the second panel of Table 2: the location fundamentals, the entry costs, and the multipliers on the trade

Table D.2: Comparing the Non-Zero Mean Estimations with the Baseline

	Dispersion		Corr. w. Prod.		Mean	
	(1) output	(2) labor	(3) output	(4) labor	(5) output	(6) labor
Dispersion, output, baseline	0.786*** (0.071)					
Dispersion, labor, baseline		31.165** (14.208)				
Corr. w. prod., output, baseline			0.142 (0.130)			
Corr. w. prod., output,baseline				0.314* (0.166)		
Dispersion, output, non-zero mean					6.830*** (0.205)	
Dispersion, labor, non-zero mean						0.324*** (0.004)
Constant	0.050*** (0.008)	-25.262 (16.953)	-0.292*** (0.033)	0.015 (0.018)	-0.574*** (0.027)	0.712*** (0.178)
N	237	237	237	237	237	237
Adj.R-squared	0.337	0.016	0.001	0.011	0.825	0.967

Notes: the table compares the estimated parameters in the baseline estimation to those estimated under the non-zero mean assumptions. Standard errors are in parentheses. ***: significant at the 1% level; *: significant at the 5% level; *: significant at the 10% level. The dependent variables of the columns are those estimated under the non-zero mean assumptions.

and migration matrices. To compute the semi-elasticities of frictions, we move the σ, ρ, or μ parameters toward zero by 0.01. In the case of σ, this amounts to a reduction of 0.01. In the cases of ρ and μ, this exercise reduces their absolute values by 0.01. We then compare the key endogenous variables in the counterfactual simulations to those under the baseline estimation. Table D.3 summarizes the results. The first row of the table repeats the baseline results reported in the main text for reference.

Different from the baseline results, under the non-zero mean assumption, reducing the frictions' dispersion parameter could reduce aggregate welfare. The first column of Table D.3 highlights these results. While the semi-elasticity of σ is -3.19 in the baseline result, it turns positive to 0.42 under the non-zero mean assumption. Further decomposing the positive semi-elasticity indicates that the reversal comes from the output friction with a semi-elasticity of 1.56. In contrast, the semi-elasticity in labor friction is still negative at -1.17.

The seemingly puzzling effect of dispersion comes from rich interactions between σ and

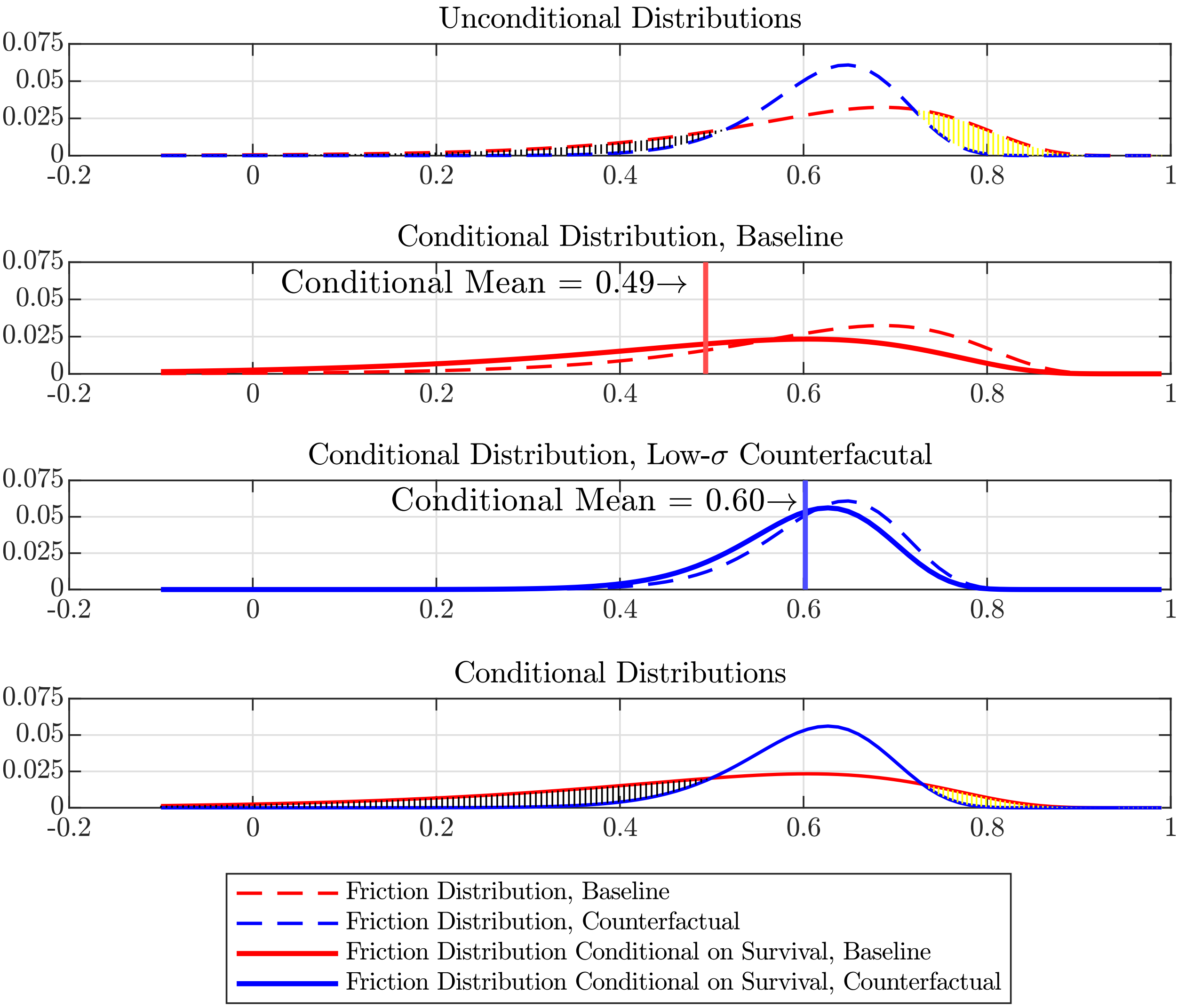


Figure D.1: The Conditional Mean of Frictions

Notes: the first panel plots the unconditional probability density function of the output friction, τ_y, such that $\log(1-\tau_y)$ follows a normal distribution with mean μ_y and standard deviation σ_y. The two dashed lines have the same μ_y. The red dashed line has a baseline level of σ_y, and the blue dashed line has a counterfactual level of σ_y lower than the baseline. The second panel compares the friction distribution conditional on survival (the red solid line) to the unconditional one under the baseline scenario. Firms with higher frictions are less likely to survive, and therefore the solid density functions are more skewed to the left. The vertical line is the conditional mean of friction. The third panel repeats the exercise for the counterfactual distribution. Note that as selection drives out the firms with higher friction, the conditional mean of the distribution with a lower σ_y (0.60) is higher than that from the baseline distribution (0.49). The last panel compares the two conditional distributions between the baseline and the counterfactual cases. Compared to the unconditional distributions reported in the first panel, the yellow shaded area shrinks when the firm faces reduced friction in the counterfactual. At the same time, the dark shaded area, in which the firm faces higher frictions in the counterfactual, expands due to the selection effect.

Table D.3: The Semi-Elasticities under the Non-Zero Mean Assumptions

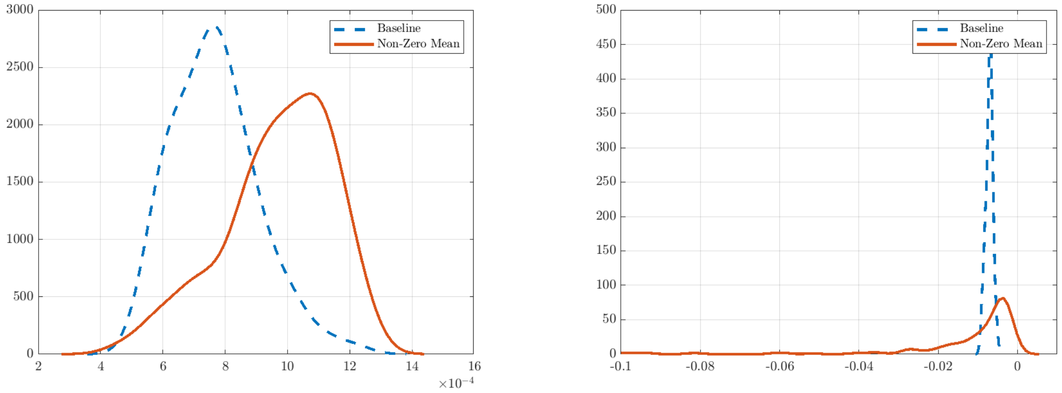
	Aggregate Welfare			Firm Entry		
	σ	ρ	μ	σ	ρ	μ
Baseline	-3.19	-0.84	-	-2.47	-0.49	-
Both Frictions, Non-Zero Mean	0.42	-3.35	-1.75	2.33	-1.73	-2.22
Output Frictions, Non-Zero Mean	1.56	-2.42	-1.77	3.66	-1.31	-2.40
Labor Frictions, Non-Zero Mean	-1.17	-0.83	0.02	-1.38	-0.40	0.18

Notes: the table reports the aggregate implications under the assumption of the non-zero mean. The numbers in the table are the semi-elasticities of reducing σ, ρ, or μ by 0.01. The first row repeats the baseline results for reference. The second row performs the same exercise, using the parameters estimated under the non-zero mean assumption. The third row reports the results in which we only change the parameters related to the output frictions, and in the last row, we only change the parameters related to the labor frictions.

the selection effects. In short, lowering σ could reduce welfare by increasing the mean friction conditional on the firm's survival. This effect is particularly pronounced when the unconditional mean of friction is positive. Figure D.1 illustrates these effects. The top panel plots the density function of τ_y from a hypothetical prefecture in the non-zero mean estimation, where the unconditional mean of τ_y is positive. Recall that we assume $\log(1-\tau_y)$ follows a normal distribution with mean μ_y and standard deviation σ_y. The red dashed line is the density function of τ_y, and the blue dashed line is the density of a counterfactual distribution with the same μ_y but a lower σ_y. As expected, the counterfactual distribution with a lower dispersion is more clustered around the unconditional mean, as the mass at both the right and the left tails move toward the center. Note that reducing the dispersion leads to opposite effects on the two tails. On the one hand, moving firms away from the right tail reduces the friction they face, as indicated by the yellow shades on the right. On the other hand, however, moving the firms from the left tail toward the center effectively increases their friction, as indicated by the dark shades on the left.[26]

The selection of firms subsequently amplifies the adverse effects on the left tail and, at the same time, dampens the positive effects on the right. This is because firms on the right tail are much less likely to survive than those on the left. The second panel of Figure D.1 contrasts the unconditional distribution to the one conditional on survival in the baseline

[26] The unconditional mean of a log-normal distribution is $\exp(\mu + \frac{\sigma^2}{2})$, which means that a reduction in σ typically reduces the unconditional mean, as shown in the top panel of Figure D.1. However, the selection effect we discuss later offsets the decline in the unconditional mean and leads to a higher conditional mean, as shown in the second and third panels.



(a) Output Frictions

(b) Labor Frictions

Figure D.2: The Changes in Conditional Mean of Frictions with a Lower σ

Notes: the figures show the histograms of the changes in the conditional mean of output (labor) frictions when we lower σ_y (σ_ℓ) by 0.01 in the left (right) panel. The conditional mean is computed from the friction distribution, conditional on firm survival with a non-negative profit. The blue lines use the baseline parameters, and the red lines use the parameters estimated with the non-zero mean assumption.

case. The third panel does the same for the counterfactual case with a lower σ. In both cases, the conditional distributions shift to the left, as those in the right tail are less likely to survive. However, as the firms on the right tail see a reduction in friction in the counterfactual, the selection effect could offset or reverse the welfare gains associated with lower dispersion. The two vertical lines in the second and third panels of the figure indicate the conditional means of friction in the baseline (0.49) and the counterfactual cases (0.60): indeed, the selection effect implies that reducing σ could lead to an **increase** in the mean friction conditional on survival, which in turn reduces welfare. To further highlight this mechanism, the last panel of Figure D.1 plots the two conditional distributions together. The selection effect is more prominent than the unconditional distributions reported in the first panel. The yellow-shaded area where the firms enjoy lower frictions in the counterfactual shrinks in the conditional distribution, as these firms are less likely to survive. At the same time, the dark-shaded area where the firm faces higher frictions in the counterfactual expands. Eventually, the conditional mean of frictions increases in the counterfactual with a lower σ.

The effects described above are more pronounced in the case of output frictions under the non-zero mean assumption. Figure D.2 presents the histogram of the changes in the conditional mean of frictions when we reduce σ across 237 prefectures. The red solid lines are based on the parameters estimated under the non-zero mean assumptions, and the blue dashed lines are based on the baseline case with $\mu_y = \mu_\ell = 0$. As the figure shows, the conditional mean in most prefectures increases when σ_y declines, leading to an adverse welfare impact as reported in Table D.3. The reason behind this is the estimated unconditional mean of the frictions: while only 138 prefectures have positive unconditional labor frictions, 228 have positive unconditional output frictions. The changes in conditional mean are also milder in the baseline estimation with $\mu = 0$, in which the unconditional means are close to zero.[27]

The selection effects discussed above directly apply to the extended model with fixed costs. In the baseline model without fixed costs, no selection occurs as all entering firms stay. Nevertheless, instead of exiting the market, the firms with unfavorable draws of the frictions will now operate on tiny scales in all equilibria. Quantitatively, these "zombie firms" leave negligible impacts on aggregate variables. The primary mechanism is thus similar to the one in the baseline economy.

Lastly, Table D.3 also confirms that reducing the absolute values of μ parameters typically leads to welfare improvement, especially in the case of output frictions with a semi-elasticity of -1.77. The welfare effects are expected, as reductions in μ typically lower both the unconditional and conditional frictions in all locations. The semi-elasticity of labor frictions is much smaller at 0.02 and not significantly different from zero, mostly because the estimated μ parameters for labor frictions are much closer to zero than that of the output friction: the average μ_ℓ is only 0.06, while the average μ_y is -0.46.[28]

E Data

Annual Surveys of Industrial Firms (ASIF) We use the ASIF panel data from 1998 to 2007. Our sample excludes the firms whose sales revenue is less than the wage bill or the value of intermediate inputs and those with non-positive value-added, total assets, fixed assets, or equity. We also drop the firms with fewer than 20 employees as the payroll data

[27] The unconditional mean in the baseline case is not zero when $\sigma > 0$. In particular, the unconditional mean of the labor friction is $\exp(\sigma_\ell^2/2) - 1$, and of the output friction, $1 - \exp(\sigma_y^2/2)$.

[28] The unconditional mean of output friction is $1 - \exp(\mu_y + \sigma_y^2/2)$, and therefore a lower μ_y indicates higher average τ_y.

in these firms might be subject to higher measurement errors. Prefectures with less than 500 firms are also excluded. In the 237 prefectures in our sample, the dataset contains 1.05 million firms during 1998-2007. Around 92.9 percent of the firms are private.

We use the “user costs of capital” as the proxy for capital costs. The user costs of capital are the rate of return from total assets, together with the depreciation of the fixed asset in the current year. Total assets include both fixed assets and flow of funds. The interest rate is a simple average of the lending interest rate between 2000 and 2005 reported in the World Bank Indicator. We understand that the changes in the price of capital goods would be a better measure of the capital costs. However, as we do not have access to such information, we settle with a less-than-ideal but reasonable measure of capital costs.

Economic Census Data on the total number of firms in each prefecture are collected from the *Second Economic Census* conducted by the National Bureau of Statistics in 2008. The economic census in China is conducted every five years and covers all the legal entities, establishments, and self-employed enterprises of the second and tertiary industries. Each prefecture-level city’s Bureau of Statistics publishes its Communique of the Second National Economic Census. The $No.1$ Communique provides the number of enterprise legal entities that we use as an approximation for the total number of entering firms, which is required for estimating firm-level frictions and calibration. The total number of enterprise legal entities in our sample is 3,616,432.

Population Census The population data come from the *Population Census* in 2000. The population Census in China is conducted every ten years and provides the resident population of each prefecture. Unlike the Hukou population, an individual can be considered a prefecture resident only if he/she has lived there for more than half a year. We use the number of residents in each prefecture as the indicator for the initial population distribution in 2000. To compute the aggregate stay rate to calibrate the scale of the migration costs $\bar{\lambda}$, we relied on the *One Percent Population Survey* in 2005. The survey is conducted every five years and covers around 17.05 million respondents, which took up about 1.31% of the total population in 2005. The advantage of the one-percent population survey over the population census is that it provides more detailed information about the respondent, including the current residence and his/her residence five years ago, which enables us to calculate the aggregate stay rate.

Investment Climate Survey This survey covers 12,500 firms in mainland China. Each firm was asked to report the percentage of sales by destination: within the prefecture, within the province, across provinces, or overseas. On average, 62.5 percent of the total revenue was generated from sales outside the local prefecture.

City Statistical Yearbook *City Statistical Yearbook* is an annual statistical publication containing socioeconomic data of cities at the prefecture and county levels. We use the per-capita GDP data at the prefectural level to calibrate the prefecture-specific productivity.

Input Output Tables We use the IO table published in 2002 to compute the labor share at the national level. In particular, labor share is the ratio between the total value-added and total output across all non-agriculture industries.