Reassessing the Firm Selection Hypothesis: New Evidence from Chinese Highways*

Jiawei Chen Yi Niu Matthew Shum

January 2024

Abstract

This paper re-investigates whether larger cities eliminate more low-productivity firms, the so-called *firm selection hypothesis*. We exploit a huge boom of infrastructure construction in China during 1998-2007. We find that difference in firm selection is quite apparent when we compare large cities to small cities which are not connected by controlled-access highways; however, the difference fades once we compare large cities to small ones which are connected by highways. This result suggests that market size is dictated less by geography (city boundaries), but rather by transportation costs. The estimated effects of firm selection are robust to the potential endogeneity of highway construction. Moreover, evidence for firm selection is generally absent in provinces with underdeveloped market economies, except in those sectors with the largest extent of liberalization.

Keywords: firm productivity, firm selection, market size, transportation infrastructure.

1 Introduction

In the literature in trade and spatial economics, an important hypothesis linking firm productivity to market size is that a larger market leads to tougher competition and therefore raises the productivity cutoff and eliminates more least-productive firms (Syverson, 2004a and 2004b; Melitz and Ottaviano, 2008; Behrens et al., 2017). This firm selection hypothesis provides insights into a range of topics in trade, geography, competition and efficiency (Melitz and Redding, 2014; Behrens and Robert-Nicoud, 2015; Proost and Thisse, 2019).

*Chen: University of California, Irvine (Irvine, CA, USA), jiaweic@uci.edu. Niu: Capital University of Economics and Business (Beijing, China), yiniu@cueb.edu.cn; corresponding author. Shum: Caltech (Pasadena, CA, USA), mshum@caltech.edu. We thank Yongheng Deng, Jindong Pang, Matt Turner, Jian Wang, Jia Yan, Junfu Zhang, and participants at the 2019 International Symposium on the Frontier of International Trade and Regional Science (Wuhan), the Urban Economics Association 2020 virtual conference, the Virtual Workshops on Topics in Regional and Urban Economics (Jinan-Peking University), Brookings-Jinan China Microeconomic Policy Forum 2021, the 4th International Workshop on Market Studies and Spatial Economics (Université Libre de Bruxelles & ECARES), and the 1st Summer Meeting in Urban Economics, China (Peking University) for helpful suggestions. Qinwen Deng, Wanxin Meng and Sinan Ni provided excellent research assistance. Yi Niu appreciates the funding support from the National Natural Science Foundation of China (Grant No. 71903138).

To empirically assess the firm selection hypothesis, Combes, Duranton, Gobillon, Puga, and Roux (hereafter CDGPR, 2012) introduce an influential quantile-based approach that estimates whether and to what extent larger markets left-truncate the productivity distribution more than smaller ones, while controlling for right-shift and dilation of the productivity distribution possibly induced by agglomeration economies, which the prior literature overlooked. Applying the approach to manufacturing firms in France, they find little evidence the productivity distribution in larger cities is more left-truncated than the one in smaller cities, in apparent conflict with the firm selection hypothesis. However, it is unclear whether their finding suggests a falsification of the firm selection hypothesis or similar strength of selection in all the cities within a larger, unified market (such as France).

In this paper we revisit the hypothesis, using firm-level data of manufacturing sectors from China. Our main contribution is to show that once we condition on cities' proximity to controlled-access highways (henceforth referred to as highways, analogous to interstate freeways in the United States), the results completely change. Difference in firm selection is quite apparent when we compare large cities to small cities which are *not* connected by highways: the productivity distribution of smaller cities contains a left tail of lower-productivity firms which is absent from the distribution of larger cities. However, the difference in firm selection fades once we compare large cities to small cities which are connected by highways.

This result suggests that market size is dictated not so much by geography, but rather by transportation costs. This is recognized in the trade literature, where a market is typically defined as a geographical entity such that agents face zero (or low) transportation costs of trading to locations within the entity, but face substantially higher transportation costs to reach destinations outside the entity. The presence of highways proxies reasonably for transportation costs; hence, small cities unconnected to highways are indeed small isolated markets, and less competition within small markets implies less firm selection relative to large cities. However, small cities connected to highways are plausibly part of a larger market, which explains why they exhibit no differences in firm selection relative to large cities.

Our study exploits a huge boom in infrastructure construction in China in recent decades. The aggregate length of highways was only 500 kilometers (km) in 1990, but rose quickly to 4,800 km in 1997, 53,900 km in 2007, and 161,000 km in 2020. These new highways have linked small cities to large regional cities and merged them into larger market areas. This provides a unique testing ground for us to revisit the firm selection hypothesis by examining the productivity distribution of firms in these cities before and after new highways appeared.

This paper combines two strands of literature which have previously developed somewhat distinctly. First, the literature on firm selection, which largely overlooked the role of transportation costs, has reported mixed results. On the empirical front, scant evidence of firm selection from city size has been found for manufacturing sectors in France, Japan, and Italy during the 1990s and 2000s (CDGPR, 2012; Kondo 2017; Accetturo et al. 2018), while Syverson (2004a) reports confirmatory evidence for firm selection from market size in his study of the ready-mixed concrete industry. Backus (2020), also studying the ready-mixed concrete sector, finds less evidence for firm selection; rather, the competition in local markets (measured by the density of concrete firms) improves firm productivity but does not eliminate low-productivity firms. Our paper provides an explanation which reconciles these divergent results, from the point of view of transportation costs: France, Japan and Italy all have mature and well-developed nationwide transportation infrastructures, while transportation costs for concrete are prohibitive even for small distances, leading to extremely localized and geographically segmented markets.

Our paper also sheds new light on the theoretical debate about the strength of the firm selection effect, where prior studies have similarly come to sharply different conclusions. For instance, the main model of Behrens et al. (2014) predicts constant selection across cities, while Behrens and Robert-Nicoud (2014) and Behrens et al. (2017) both predict stronger selection in larger cities. By highlighting the critical role of market characteristics—particularly the magnitude of transportation costs and the development of market economy—in determining the strength of the firm selection effect, our paper suggests a mechanism for accommodating the opposing theoretical predictions and provides a useful setup for future investigation of this topic.

Second, several papers have likewise exploited the growth in transportation infrastructure in China to empirically test economic theories, but they have not looked at the firm selection hypothesis. Looking more historically, Banerjee et al. (2020) document the relation between recent economic growth in Chinese regions to proximity to imperial-era transport hubs, finding that highways and railroads promote economic growth. However, Faber (2014) shows that increasing highway connections have *reduced* GDP growth in counties adjacent to new highways. Importantly, he uses a novel approach based on "shortest path" distances between cities to construct exogenous instruments for highway construction, which we will also use in this paper. Baum-Snow et al. (2020) report spatially contrasting effects from highways. Specifically, regional highways increase various economic outcomes, such as GDP, population, and wage, for larger "regional primate" prefectures but reduce these outcomes for smaller "hinterland" prefectures. These papers have not examined the relation between highways and firm productivity, which is

the primary focus in this paper.

Three studies use CDGPR's quantile approach and find certain evidence for the selection effect on TFP distributions. Ding and Niu (2019) report stronger selection effect in larger provinces of China, Accetturo et al. (2018) report stronger selection effect in Italian cities and provinces with greater market potential, and Arimoto et al. (2014) report stronger selection effect in industrial clusters than elsewhere of Japan. We differ from them in the following aspects. First, we provide the first evidence for the selection effect from city size using the quantile approach, while they either do not investigate city size (Ding and Niu, 2019; Arimoto et al., 2014) or do not find any evidence from city size (Accetturo et al., 2018). Second, as the obstacles to trade that Ding and Niu (2019) and Accetturo et al. (2018) focus on are either unobservable (informal trade barriers between provinces in Ding and Niu) or fixed over time (Euclidean distance in Accetturo et al.), they are unable to explore the dynamics of firm selection with evolving trade costs, proxied by highway construction, like this paper. Third, the selection effect found by Arimoto et al. (2014) is driven by competition in the input market, whereas our paper focuses on competition in the product market.

2 Controlled-Access Highways and Transportation Costs

Controlled-access highways are vitally important for domestic trade within China; the share of domestically traded goods delivered via trucks has surpassed 70% since 1985, and reached 78% in 1998, the beginning of our study period.[1] Highway infrastructure in China has grown phenomenally within the past three decades. From 1990 to 2019, the length of all types of highways increased over six-fold, from 0.74 to 4.70 million km, with the fastest growth being in controlled-access highways, which are the largest roads in terms of both lanes and traffic.[2] Figure 1 shows that, from 1990 to 2019, the length of controlled-access highways increased around 300 times, from 500 to 149,600 km, with an annual growth rate of 21.7%. The increase in controlled-access highways was especially pronounced during our study period, growing from 4,800 km in

[1] Source: *China Statistical Yearbook*, various years.

[2] Highways in China, the so-called *dengji gonglu*, are divided into five classes, controlled-access highways and Classes 1-4. The controlled-access highways have divided lanes and access control, and are only used by automobiles. The designed average daily traffic (ADT) is 25,000 – 55,000 automobiles for four-lane controlled-access highways, 45,000 – 80,000 for the six-lane, and 60,000 – 100,000 for the eight-lane. The first-class highways have divided lanes but not necessarily access control. Their designed ADT is 15,000 – 30,000 automobiles for four-lane highways, and 25,000 – 55,000 for six-lane highways. The second-class highways have two lanes and designed ADT of 3,000 – 7,500. The third-class highways have also two lanes but a lower ADT of 1,000 – 4,000. The fourth-class highways have either two lanes or a single lane, with designed ADT below 1,500 for the two-lane, and below 200 for the single-lane.

1997 to 53,900 km in 2007, a growth rate of 27.4% per year.[3]

[Figure 1 here]

Controlled-access highways differ from other roads mainly because they are designed exclusively for high-speed, unhindered vehicular traffic. In particular, they have no traffic signals, intersections or property access, do not allow pedestrians, non-motorized vehicles or farm machinery, have high-quality road surfaces, and allow vehicles to travel at high speeds up to 120 kilometers per hour (kph).[4] Not surprisingly, controlled-access highways have dramatically reduced travel times within China: during the 1980s, for instance, driving from Beijing to Tianjin (about 160 km) took approximately 7 hours, and it took almost one day to travel the 300 km to Shijiazhuang, a nearby provincial capital city. Two new controlled-access highways built in 1993 and 1994 cut the travel time to 1.5 hours and 3 hours, respectively. Similarly, the new controlled-access highways reduced drive times between Shanghai and Nanjing (1996; from 6-7 to less than 3 hours), Shenyang and Panjin (2000; from 3 hours to 1), and Changsha and Yongzhou (2003; from 6 to 3 hours).[5]

Controlled-access highways significantly lower the costs of shipping goods for a number of reasons. First, the higher driving speeds reduce time costs and hence wages paid to drivers. Second, the higher speed limit and fewer turns and brakes together increase vehicular fuel efficiency and reduce fuel costs per kilometer, as well as repair and maintenance costs. Third, the improved surfaces reduce transit damage (particularly for fragile items) and permit heavier loading of trucks which leads to larger scale economies in transportation. Fourth, accident rates are lowered largely due to the access control.

For instance, in a 1998 survey on the controlled-access highway connecting Shijiazhuang with Beijing, Wu (2005) found that it reduced the costs of repair and maintenance by 77% and drivers' wages by 72%, as compared to a second-class highway (National Highway #107) which ran parallel to the controlled-access highway. Jia et al. (2004) find that, for a medium-sized truck running at the speed of 100 km/h, the Changping Highway in Liaoning province reduces fuel consumption by 30% compared to a parallel national highway #102.

[3]For comparison, in 2018, the length of controlled-access highways is 77,960 km in US and 78,097 km in EU.

[4]For comparison, the speed limit is 60-100 kph for Class 1 highways, 40-80 kph for Class 2, 30-60 kph for Class 3, and 20-60 kph for Class 4.

[5]Sources are as follows. Beijing-Tianjin: http://www.xinhuanet.com/politics/2018-08/06/c_1123270958.htm; Beijing-Shijiazhuang: http://m.xinhuanet.com/2018-07/13/c_1123121492.htm; Shanghai-Nanjing: http://58.213.139.243:8088/imgpath/zz3/1996-4/D1/D1.html; Shenyang-Panjin: http://news.sina.com.cn/china/2000-09-15/127175.html; Changsha-Yongzhou: http://news.sina.com.cn/o/2003-12-27/08271443738s.shtml; all accessed on November 1, 2020.

3 Theory and Empirical Approach

3.1 Theory

In this section, we provide theoretical background to show that, in the presence of sizeable transportation costs, large cities generate stronger firm selection than small cities. Furthermore, as transportation costs decline, the differences in the strength of firm selection in big and small cities converge, as smaller transportation costs essentially imply that the cities are part of a single market. The discussion is based on models from Melitz and Ottaviano (2008) and CDGPR (2012).

The starting point for our theoretical discussion is a model of intra-provincial trade. Such a model is warranted as the Chinese commercial market may be strongly segmented at the provincial level, due to historical and political reasons. Under China's incremental reform and decentralization policies, the performance and promotions of local officials are evaluated largely based on local economic growth, so provincial governments set barriers to domestic trade in order to protect local firms from regional competition and thereby to maximize their local GDP and fiscal revenues (Young, 2000). Such barriers operate not via formal taxes or tariffs, but are usually informal, indirect and internal. Examples include sale permits, internally announced preferential policies towards local suppliers, stricter inspections and extra charges on regional imports, and so on.[6] The strength of the inter-provincial trade barriers around our study period, 1998-2007, however, remains unclear. One strand of the literature implies quite strong trade barriers between provinces (Young, 2000; Poncet, 2003; Poncet, 2005); among the rest, Poncet (2005)'s estimates suggest that China's inter-provincial trade barriers during 1992-1997 were roughly the same magnitude as those within the European Union countries or between US and Canada in the 1990s, and have been growing over time. Another strand of the literature believes the barriers should be weak (Naughton 2003; Holz, 2009; Xing and Li, 2011; Xu and Fan, 2012); among them, Xing and Li (2011)'s estimates of the inter-provincial trade barriers during 2003-2005 are as low as those of the inter-state trade barriers during the 1990s in the US. In the following model, we will discuss both situations with significant and negligible inter-provincial trade barriers, respectively.

There are I cities divided into U provinces ($I > U$), and the population of city i in province u is denoted by L_u^i. An individual consumer's utility is given by

$$U = q^0 + \alpha \int_{k\in\Omega} q^k dk - \frac{1}{2}\gamma \int_{k\in\Omega} \left(q^k\right)^2 dk - \frac{1}{2}\eta \left(\int_{k\in\Omega} q^k dk\right)^2 \quad (1)$$

[6] Gilley (2001) and Li et al. (2004) have more detailed descriptions on specific forms of these trade barriers.

where q^0 is the individual's consumption of a homogeneous numeraire good, q^k is the consumption of a variety k of differentiated goods from a set Ω and $\alpha, \gamma, \eta > 0$ are taste parameters. Under this setting, there exists a choke price $\overline{p}$, so that any varieties of differentiated goods more expensive than $\overline{p}$ will not be consumed.

To produce one variety of a differentiated good, a firm incurs a sunk cost f_E, and then randomly draws the marginal cost c from a common distribution with probability density function (PDF) $g(c)$ and cumulative density function (CDF) $G(c)$ (each firm only produces one variety due to increasing returns). Any firm drawing a marginal cost exceeding the choke price $\overline{p}$ earns no operational profits and hence exits the market. The numeraire is produced competitively without any sunk cost.

Delivering differentiated goods to consumers involves two types of iceberg trade costs: transportation costs between cities and interprovincial trade barriers. To deliver one unit of a differentiated good to another city within the same province, the number of units a firm needs to ship is $\tau_1 = 1 + \kappa$, where $\kappa \geq 0$ represents transportation costs. To have one unit arrive at a city in another province, the number of units the firm needs to ship is $\tau_2 = 1 + \kappa + \lambda$, where $\lambda \geq 0$ represents provincial trade barriers. Shipping differentiated goods inside a city or shipping the numeraire anywhere is assumed free.

Under free entry, the ex-ante expected profits must be driven to zero, so for any firm with marginal cost c located in city i and province u, we have

$$\frac{L_u^i}{4\gamma}\int_0^{\overline{p}_u^i}\left(\overline{p}_u^i - c\right)^2 g\left(c\right)dc + \sum_{j\neq i}\frac{L_u^j}{4\gamma}\int_0^{\frac{\overline{p}_u^j}{\tau_1}}\left(\overline{p}_u^j - \tau_1 c\right)^2 g\left(c\right)dc + \sum_{v\neq u}\sum_k \frac{L_v^k}{4\gamma}\int_0^{\frac{\overline{p}_v^k}{\tau_2}}\left(\overline{p}_v^k - \tau_2 c\right)^2 g(c)dc = f_E \tag{2}$$

where the three components on the left-hand side of the equation represent operational profits the firm earns from selling in the local city, selling to other cities within province u, and selling beyond province u, respectively. In city i of province u, the share of exiting firms is $S_u^i = 1 - G(\overline{p}_u^i)$, where the cutoff $\overline{p}_u^i$ is endogenously determined and depends on both city size (if $\kappa > 0$) and provincial size (if $\lambda > 0$).

If inter-provincial trade barriers are high ($\lambda > 0$), a large city does not necessarily have a lower value of $\overline{p}_u^i$ than a small city in a different province, for the provincial size also affects $\overline{p}_u^i$. But within a province, larger cities have a lower value of $\overline{p}_u^i$ and hence a higher S_u^i: that is, larger cities eliminate more low-productivity firms. This is the *firm selection hypothesis* (see Online Appendix A for the proof):

Hypothesis (Firm selection): If $\lambda > 0$, then for any two cities i and j in a province u such that

$L_u^i > L_u^j$,

1. Under high trade costs between cities ($\kappa > 0$), the larger city left-truncates a larger share of firms: $S_u^i > S_u^j$;

2. Under low trade costs between cities ($\kappa = 0$), the large and small cities left-truncate the same share of firms: $S_u^i = S_u^j$.

If inter-provincial trade barriers are low ($\lambda = 0$), then the above model reduces to CDGPR (2012)'s model. It predicts that, for any two cities m and n such that $L^m > L^n$, whether they are in the same province or not, $S^m > S^n$ if $\kappa > 0$, and $S^m = S^n$ if $\kappa = 0$. Since whether inter-provincial trade barriers are high or low during our study period remains controversial in the literature, to investigate the hypothesis under both scenarios, we will compare large and small cities both in the same province and from multiple provinces.

3.2 Empirical Approach

To test whether a larger city left-truncates a larger share of firms (the Firm Selection Hypothesis), we use the quantile approach developed by CDGPR (2012), which estimates the selection effect while accommodating agglomeration economies arising from city size. We describe this briefly here.

Denote a firm's log productivity as $\phi = \log(1/c)$. Given the underlying CDF of marginal cost $G(c)$, the underlying CDF of log productivity is $\tilde{F}(\phi) = 1 - G(e^{-\phi})$. As a larger city may have stronger agglomeration economies that increase the productivity of all active firms in the city, we can express the observed CDF of log productivity in city i at province u as

$$F_u^i(\phi) = \max\left\{0, \frac{\tilde{F}_u\left(\frac{\phi - A^i}{D^i}\right) - S_u^i}{1 - S_u^i}\right\}$$

where A^i represents the extent that agglomeration right-shifts the distribution; D^i is a dilation term which indicates whether more productive firms benefit more, the same, or less from agglomeration, depending on whether D^i is greater than, equal to, or less than one.

Consider city i and j such that $L^i > L^j$. The CDF's of firm productivity in the large city i and the small city j are related by the following transformation:

$$F^i(\phi) = \max\left\{0, \frac{F^j\left(\frac{\phi - A}{D}\right) - S}{1 - S}\right\} \tag{3}$$

where $S = \left(S^i - S^j\right) / \left(1 - S^j\right)$, $A = A^i - DA^j$, and $D = D^i / D^j$; these three parameters represent, respectively, the strength of selection, agglomeration and dilation in the large city relative to the small one. We follow CDGPR (2012) and Gobillon and Roux (2010) to estimate the parameters of S, A and D, and use the bootstrap method to estimate their standard errors.[7]

To test the Firm Selection Hypothesis, we will fit equation (3) to the estimated firm productivity distributions in large vs. small cities. When the two groups of cities are not connected by highways and are therefore segmented, we expect a stronger firm selection effect in large cities, i.e., $\hat{S} > 0$, as illustrated by Panel A of Figure 2. When the two groups of cities are connected by highways, however, we expect the gap in firm selection to narrow, i.e., $\hat{S} = 0$, as illustrated by Panel B of Figure 2, because small and large cities connected by highways are essentially the same market and hence eliminate the same proportion of firms.

[Figure 2 here]

3.3 Focus on Market Economy

As firm selection requires market competition, in the empirical investigations we primarily use selected provinces that have relatively strong market forces during China's market transition.

Under China's piecemeal approach to market reform, some regions have been the first to experience these reforms, while some have lagged behind. For instance, in order to welcome foreign investment and expand foreign trade, early special policies and designations (e.g. lower tax rate, more authority to approve foreign investment, special economic zone, etc.) are granted to, in sequential order: Guangdong and Fujian provinces around 1980, 14 coastal open cities in 1984, Hainan province in 1988, and most prefecture cities along the Yangtze river and the national borders in 1992 (Qian, 2000); the privatization of SOEs was first begun in the provinces of Shandong, Guangdong and Sichuan, and then gradually extended to other provinces (Qian, 2000). Some eastern provinces were the first to accommodate many non-state-owned enterprises, including town and village enterprises, foreign-invested enterprises and private enterprises (Fujita and Hu, 2001), and they also established a sophisticated market system much earlier than other regions involving, for example, better integration of developed product markets, factor markets, and intermediary markets, as well as a more efficient legal system (Hao and Wei, 2010).

We evaluate the development of market economies by province in 1998 according to the following indicators. The first three include the share of SOEs in the number of above-scale industrial firms, the share of SOEs in the employment of the above-scale industrial firms, and

[7] We use the code published at http://diegopuga.org/data/selectagg/ for the estimation.

the share of SOEs in the total value added of the industrial sectors. The fourth indicator is the marketization index developed by Wang et al. (2009), who evaluate and rank the degree of marketization by province from five aspects: the relationship between government and market, the growth of non-state economies, the development of product market, the development of factor market, and the development of intermediary organizations as well as legal environment. As shown in Table 1, the five provinces of Zhejiang, Jiangsu, Fujian, Guangdong and Shandong exhibit much better developed market economies than the rest of China, as their rankings are consistently higher than most of other regions: they are the top five in the first two indicators, and they remain top seven and top six in the third and fourth indicators, respectively. In the last column, the sum of ranks, the five provinces are significantly higher than others. Therefore, we primarily use the five provinces with well-developed market economies to test the firm selection hypothesis, and will also investigate other provinces for comparison.

[Table 1 here]

4 Data and Variables

This paper primarily uses three datasets on the firm-level characteristics, the population size of cities, and GIS data on locations and routes of controlled-access highways. We spent substantial effort to reduce potential errors in these data, as data noise could easily blur the estimated selection effect by the quantile approach (Ding and Niu, 2019).

4.1 Firm-Level Data

The firm-level dataset derives from the Annual Survey of Industrial Firms (ASIF) during 1998-2007 conducted by China's National Bureau of Statistics (NBS). Included in the data are the so-called "above-scale" firms, which consist of all state-owned enterprises (SOEs) and those non-state-owned firms with sales exceeding 5 million RMB in the year.[8] We drop those in the sectors of mining and public utilities, and retain firms in the manufacturing sectors for this research, which make up 90.6% of all above-scale firms in 1998 and 93.0% in 2007. A rich set of firm-level variables is available, including industrial output, value added, employment, capital, input, code of administrative division, industry code, ownership and many kinds of detailed financial

[8]The above-scale firms account for approximately 27% of all firms in 1998 according to our estimate, as discussed in Online Appendix B. Missing small firms, as shown in Ding and Niu (2019), will lead to an underestimation of the selection effect given a positive correlation between firm size and firm productivity.

information. Online Appendix B documents how we checked and processed this data before using it in this paper.

4.2 Estimating Firm Productivity

Using the above ASIF data, we measure firm productivity by estimating total factor productivity (TFP) at the firm-year level. Assuming that firms' production function takes a Cobb-Douglas (log-linear) form, we obtain the following estimating equation:

$$\log(V_{it}) = \beta_C + \beta_K \log(K_{it}) + \beta_L \log(L_{it}) + \mu_{it} \tag{4}$$

where L_{it} is employment of firm i in year t, V_{it} is value added, K_{it} is capital input. The ASIF data reports the employment at the end of the year and the average employment over the entire year, and we use the latter as L_{it}. Both value added and capital are measured after deflation, as described in Online Appendix B. Using the approach developed by Ackerberg et al. (2015), we estimate function (4) separately by each of the 28 2-digit sectors, and thereby measure the TFP of firm i in year t by $\widehat{\mu}_{it} = \log(V_{it}) - \widehat{\beta}_C - \widehat{\beta}_K \log(K_{it}) - \widehat{\beta}_L \log(L_{it})$. Online Appendix C discusses the estimation of function (4) and displays the estimated results.

Finally, we take a two-year average of TFP for each firm; for instance, the TFP of firm i during 1998-1999 is computed as $\widehat{\mu}_{i,1998-1999} = (\widehat{\mu}_{i,1998} + \widehat{\mu}_{i,1999})/2$. Averaging across two years helps to average out year-by-year noise in TFP estimates and improves the precision of the firm selection estimates.

4.3 Controlled-Access Highways

We downloaded the GIS (geographic information system) data of China's controlled-access highways in 1999 that were used in Baum-Snow et al. (2017) and Baum-Snow et al. (2020). Then we checked for errors in the GIS data and made corrections according to multiple materials, including the Yearbook of China Transportation & Communications for various years, online news on the completion of corresponding highway segments, provincial gazetteers, etc. The GIS data after our corrections indicate that the total length of highways at the end of 1999 is 11564 km, very close to the 11605 km reported by NBS. See Online Appendix D for details.

4.4 Measuring City Size

In order to utilize the quantile estimating equation (3), we need to construct a measure of city size in China. It involves two considerations: (i) how to measure the spatial scope of a city, and (ii) which variable (e.g., population or employment) is used to measure city size.

For the first issue, while a city is usually considered as a unified labor market (Duranton, 2015), China has not delineated any boundaries of local labor markets such as metropolitan statistical areas in the US and employment areas in France identified based on commuting patterns. Fortunately, the urban area of each administratively designated *shi* (also known as administratively designated city) or administratively designated *xian* (also known as administratively designated county and county equivalent) roughly approximates an integrated labor market (Chan, 2007; Li and Mykhnenko, 2018; Chen et al., 2023). Therefore, we use the urban area in each *shi* or *xian* as the spatial scope of each city, and we are able to identify the *shi* or *xian* a firm is located in according to the county-level code from the ASIF data.[9] Using the data of urban land and night-light, we also found several cases where the major urban areas in a few adjacent *shi* or *xian* expanded across their administrative boundaries and became contiguous. In these cases, their labor markets might be highly integrated, and we treat such adjacent *shi* or *xian* as the same city (see Online Appendix E for details).

For the second issue, we prefer using the urban part of the permanent resident population (PRP) in the years of national population censuses to measure city size. During a population census, PRP is counted based on door-to-door visits to all households and the actual residence of at least six months in the year, while PRP in other years are estimated based on various materials such as one-thousandth population survey, *hukou* population (registered population) suffers from large differences between the actual residence and the registered residence, and complete employment data are not available[10]. Therefore, in our main study period of 1998-1999, we measure the size of each city by using the urban part of PRP in 2000 from the 5^{th} national population census in each *shi* or *xian* (referred to as city in the remainder of this paper).

We define a city as an entity with population exceeding 10,000 persons, and thereby identify

[9] In 2000, there are 659 *shi*, including 259 prefecture-level cities and 400 county-level cities, and 1622 *xian*, including 1503 counties, 116 autonomous counties and 3 banners (data source: https://www.xzqh.org/html/show/cn/2000.html). Each prefecture-level city consists of at least one city district. Each city district, county-level city or *xian* corresponds to a unique county-level code, which is the first six digits of the administrative division code.

[10] Since 1998, China Urban Statistical Yearbook reports *danwei* employment (*danwei congye renyuan*) instead of total employment by *shi*. The *danwei* employment primarily involves employment in governments as well as government-sponsored institutions, organizations and enterprises. The *danwei* employment in all prefecture regions that cover almost entire China accounted for only 12.1% of their total population in 2000, when the employment to population ratio in China was 74% according to the World Bank.

2208 cities in China, with an average population of 207,526. Among these, 426 are located in the five provinces with well-developed market economy, and these have an average population of 350,432, as reported in Table E2 in the online appendix.

4.5 Grouping Large and Small Cities

Using the above data of city size and highways, we will examine the prevalence of the firm selection effect in two types of scenarios, corresponding to 1) high transportation costs between large and small cities, and 2) low transportation costs between large and small cities.

For the first scenario, we compare small cities located far from highways (S1 in Figure 3) to all large cities (B1 and B2 in Figure 3) which can be (and typically are) located near highways. In this way, the small and large cities are segmented in terms of transportation costs, and therefore the large cities credibly belong to a larger market vis-à-vis the small cities.[11] This analysis is only available for the period of 1998-1999, because by 2007 practically all cities are located near a highway.

[Figure 3 here]

Figure 4 displays our grouping results for the first scenario, using Zhejiang province as an example. Setting the city-size population cutoff at 500,000, we first identify 35 cities shaded in blue that are smaller than the cutoff, do not have any highway within 50 km from their boundaries at the end of 1999, and do not share any boundary with those cities larger than 1 million population. In our benchmark analyses for Zhejiang province during 1998-1999, firms located in these 35 cities will be defined as the firms in small cities. We also identify 20 cities shaded in red that are larger than all the small cities. There are additional cities, marked in white which, while smaller than the cutoff, nevertheless either already had highways within 50 km from their boundaries, or shared boundaries with cities above 1 million. These latter cities will be excluded from the main analysis in Table 2-4, but we will examine them in Table 6 below.

[Figure 4 here]

For the second scenario, we compare small and large cities among cities located close to highways, and this analysis will be implemented for both periods of 1998-1999 and 2006-2007. In most of our results, we use mono-establishment firms that account for 96.9% of all firms, because firms with multiple establishments might produce in multiple cities.

[11] We did not compare small and large cities both located far from highways as such large cities (B1 in Figure 3) are too few.

5 Main Results: Large and Small Cities without Highway Connections, 1998-1999

5.1 TFP Distributions

To start, we present smoothed TFP distributions for small and large cities in the five provinces during 1998-1999 in Figure 5. Panel A compares the TFP distributions between large and small cities which do not have a highway connection between the two groups of cities (the first scenario). Panel B compares the TFP distributions for large and small cities which were all connected by controlled-access highways (the second scenario), i.e., highways exist inside each city.[12]

[Figure 5 here]

Under the firm selection hypothesis, the TFP distribution for small cities should be more "left truncated" compared to the distribution from large cities. We see some evidence of this in Panel A of Figure 5: at low productivity levels, the distributions of large cities (in solid line) have visibly thinner left tails than the small cities (dashed line). In comparison, this pattern disappears in Panel B, where the left tails of the two distributions almost become coincident.

This visual evidence clearly suggests that the presence of a highway may modulate the strength of firm selection effects. Indeed, highway access appears to accompany higher firm exit rates. For the five provinces, in cities without highway access in 1998, 57.4% of incumbent firms existing in 1999 had exited by 2005, by which point all these cities gained highway access at least within 50 km. In comparison, only 52.3% of incumbent firms exited between 1999-2005 in cities which already had highway access in 1998. The firms who exited are generally less productive: their median TFP is 30% lower than the median of the surviving non-exiting firms. We next turn to a more rigorous test using the quantile approach described earlier.

5.2 Benchmark Results of the Quantile Model

In Table 2, we pool the five provinces together, and estimate the parameters in the quantile estimating equation (3) under different cutoffs of city size during 1998-1999. We begin with preliminary specifications which have only the selection effect (columns 1-2) and both selection and agglomeration (columns 3-5) to explain the differences in TFP distributions between the large and the small cities; finally, in columns 6-9, we estimate the full specification which allows

[12] When plotting the TFP distributions or estimating the quantile model, we always trim 1% firm observations on each side of the distribution to remove outliers, which is consistent with CDGPR (2012).

for all three effects: selection, agglomeration and dilation. The model fits the data well, as the pseudo-R^2 in column 9 remains 0.98 under the different cutoffs.

Table 2 presents strong evidence for firm selection. When we use different cutoffs of city size, the estimated values of S remain positive and statistically significant (column 6); $\hat{S}$ ranges between 2.93% and 3.47%, suggesting that the group of small cities need to left-truncate 2.93-3.47% of their incumbent firms in order to achieve the same strength of selection as the large cities. This formalizes the earlier visual evidence from the graphs in Figure 5.

[Table 2 here]

In Table 3, we estimate the parameters in equation (3) by each of the five provinces, using 0.5 million population as the cutoff of city size. The results remain consistent with the firm selection hypothesis, similar to Table 2. $\hat{S}$ ranges between 2.20% and 5.54%; while the reduction in sample size enlarges the standard errors in Table 3, the estimated values of S are still significant at the 5% level.

[Table 3 here]

The predominantly confirmatory evidence for firm selection in Tables 2-3 contrasts sharply with CDGPR (2012), who find no evidence of selection in French cities during 1994-2002 in any of the 15 manufacturing sectors. However, both sets of results, despite their big differences, could be inherently consistent with the same theory sketched above. France is a highly developed country with an extensive highway network in place since around 1980; hence, French cities regardless of size are well-integrated into a unitary market where shipping costs between cities are low, and hence cities differing in size may nevertheless exhibit quite similar magnitudes of firm selection. However, as a developing country, China's national highway system came much later: for example, out of the 2,208 cities in 1999, 83.1% do not have any highway inside their boundaries, and 54.5% do not have any highway within 50 km from their boundaries. For large cities not connected to small cities by highways, the large transportation costs between them imply that they operate as segmented markets, which implies more firm selection in the larger cities, as per the Firm Selection Hypothesis and the theoretical discussion in Section 3 above. Transportation costs are capable of explaining the divergent results between France and China.

Regarding the other parameters, the estimated coefficients of A (in column 7 of Table 2 and column 2 of Table 3) are all positive and mostly statistically significant, suggesting that large cities have right-shifted their TFP distributions relative to small cities, consistent with the hypothesis of

agglomeration economies in the form of input sharing, labor pooling and knowledge spillovers.[13] Under the cutoff of 0.5 million population in Table 2, for instance, the right-shift effect alone increases the average TFP by $e^{0.2088} - 1$, or 23.2%. While $\hat{D}$ is often lower than one, only under the cutoff of the 75^{th} percentile in Table 2 is it statistically significantly different from one. It suggests larger cities right-shift different quantiles of the TFP distribution to a similar extent.

Allowing for right-shift and dilation may also capture other factors that influence TFP distributions. Larger cities or markets may attract more productive firms or entrepreneurs (Behrens, et al., 2014; Gaubert, 2018), though this effect of firm sorting may not be strong given relatively low rates of firm relocation: 1.9% of establishments relocated across American counties during 2009-2013 (Rupasingha and Marre, 2020); 4.7% of establishments relocated across French employment areas during 1993-1996, and only 2%-4% in most manufacturing sectors (Duranton and Puga, 2001). Other factors that can be captured by right-shift and dilation include natural advantages (e.g., nice weather) that simultaneously enlarge city size and increase TFP of local firms, as well as special economic zones which are disproportionately located in large cities and which boost TFP (Lu et al., 2019).

5.3 Controlling for Endogeneity in Timing and Location of Highways

The estimated results in Tables 2-3 may be plagued by an endogeneity problem as the timing of the construction of highways connecting cities may not be random. Governments, particularly at the provincial level, might favor more "promising" cities with fewer low-productivity firms and hence build highways sooner around these cities to bolster their growth. That is, policymakers may have a preference to route highways closer to cities which have more competitive market environments, where the firm selection effect is likely to be larger. If the highway access encourages city growth and makes the cities more likely to be large cities in our benchmark results, then the values of S in the equation (3) could be overestimated.

To accommodate this potential endogeneity, for the large cities that had highway access by 1999, we will only keep those that were close to highways due to largely exogenous reasons. In 1992, China's State Council approved the construction of a national highway network with seven horizontal and five vertical axes aiming to connect all provincial capitals to administrative cities (*"shi")* with non-agricultural population exceeding 500,000 persons. Highways connecting these target nodes naturally have to go through some other cities, which are, all else equal, much more likely to obtain highway access independently of economic considerations, as discussed in

[13] See Ellison et al. (2010).

Faber (2014). Given these considerations, we consider the following criteria for identifying large cities that had highway access by 1999 due to exogenous reasons: among the large cities that had highways within 50 km from their boundaries, we only keep those that (*i*) were not the target nodes in China's 1992 highway plan, and (*ii*) were within 50 km from a hypothetically efficient highway network connecting the target nodes, which is the Euclidean spanning tree suggested by Faber (2014).

Using these criteria, we construct the groups of large cities for which highway access is arguably exogenous. The large cities that did not have any highway within 50 km by 1999 are still kept. The grouping of small cities is always the same as in Tables 2-3 as none of them had any highway within 50 km in 1999.

[Table 4 here]

Table 4 displays the estimated results using the new group of large cities based on the above strategy. Overall, we find that the estimated selection effects are similar to Tables 2-3. The five provinces together, as well as the individual provinces of Jiangsu, Guangdong and Fujian still exhibit positive and statistically significant values of $\hat{S}$, which is strong evidence for firm selection. While the $\hat{S}$ for Shandong, Zhejiang, Guangdong and the five provinces together become more or less lower than those in Tables 2-3, the $\hat{S}$ for Fujian and Jiangsu become higher. In sum, we conclude that the evidence for selection found in this paper is quite robust even after accommodating potential endogeneity in the timing and location of highway construction.

5.4 Alternative Specifications

We investigate alternative specifications to assess the robustness of our findings and conduct further investigations for 1998-1999, pooling the five provinces together and using 0.5 million population as the cutoff of city size. As the above results use mono-establishment firms, we investigate to what extent using all firms changes the estimation results. The first row of Table 5 indicates that $\hat{S}$ equals 0.0325, only slightly lower than the benchmark estimate of 0.0347, and remains highly significant.

[Table 5 here]

The reasons why small cities in the benchmark results have less selection may be not only that they are segmented from large domestic cities, but also that they face higher trade costs with foreign markets because of, for instance, higher shipping costs to ports due to the lack of

highway access. This effect should not be overlooked because in 1999, 34.6% of the above-scale firms in the five provinces export, and their export values account for 24.8% of total sales of all the above-scale firms. To accommodate this effect, we drop all exporting firms (those which report positive export volume) and re-estimate the benchmark model. As displayed in the second row of Table 5, $\hat{S}$ still remains positive and statistically significant, though its value shrinks to 0.0209.

If products are very costly to ship (e.g., ready-mixed concrete), trade costs between cities (κ in Section 3.1) might still be high even if these cities are connected by highways, and therefore larger cities should have stronger selection effect. To test this hypothesis, we first measure the dollar value per kilogram shipment (DVPKS) for each 3-digit sector according to the 2002 Commodity Flow Survey in the United States (see Online Appendix F for details). Then we use firms in the 3-digit sectors whose DVPKS is very low (i.e., the shipping costs of their products should be very high), and compare the TFP distributions of big vs. small cities that all have highway access. In the 3^{rd} and 4^{th} rows of Table 5, we find that, for sectors with DVPKS < 2, $\hat{S}$ equals 0.0124 and is significant at the 10% level; for sectors with even higher shipping costs (DVPKS < 1), $\hat{S}$ increases to 0.0130 and becomes significant at the 1% level. These results are consistent with our hypothesis and imply again that market size is mainly dictated by transportation costs.

For cities with similar size, the group of cities connected by highway should have larger market size and hence stronger firm selection than the other group of cities not connected by highway. While it is impossible to find two cities of exactly the same size, we can choose cities within a certain range of size so that their population do not differ much, and the differences in firm selection between the two groups of cities should be primarily attributed to highway. We first select cities between 0.25-0.5 million population, and compare those with highway access (46 cities, the mean size 0.37 million and the median 0.36 million) versus those without highway within 50 kilometers (35 cities, the mean size 0.34 million and the median 0.33 million). The estimated results in the 5^{th} row of Table 5 indicate that $\hat{S}$ equals 0.0229 and significant at the 1% level. In the 6^{th} row, similarly, we select cities between 0.5-0.75 million population, compare those with highway access (8 cities, the mean size 0.59 million and the median 0.57 million) versus those without highway within 50 kilometers (9 cities, the mean size 0.58 million and the median 0.56 million), and find that $\hat{S}$ equals 0.0232 and significant at the 5% level.

In the above analyses, we pool firms in all manufacturing sectors, as firm observations are insufficient in most of the 2-digit sectors. Nevertheless, we still find five subsectors that have reasonably sufficient observations, and we estimate the selection effect in each of them. The esti-

mated results in Online Appendix G suggest that the evidence for firm selection is still present.

6 Additional Results

6.1 A Placebo Test: Cities Proximate to Highways in Both 1998-99 and 2006-07

The main results thus far in the paper are derived from comparing large vs. small cities which were not connected by highways in 1998-99. As a placebo test, we now compare large vs. small cities which all had highway access. If our claim that the difference in firm selection effects between large vs. small cities are dampened by highway proximity indeed holds, then we should expect little difference in firm selection in this group of cities, in either 1998-1999 or 2006-2007. To implement this, we fit the quantile equation (3) for a group of large and small cities which all had highways within their boundaries in 1999 (for the five provinces, 25.1% of cities fell into this category) and 2007 (66.9% of cities in the five provinces included).

Indeed, the difference in firm selection effects practically disappears in this group of cities: Table 6 shows that the estimated values of $\hat{S}$ in these cities are much smaller than those reported in Tables 2-3. It is below 1% in all provinces, except for Guangdong. Further, $\hat{S}$ is no longer statistically significantly different from zero. Moreover, as the standard errors in Table 6 are actually smaller than in Tables 2-3, the lack of statistical significance is primarily due to the much smaller magnitudes of the coefficients: we have a "precisely-estimated zero".

[Table 6 here]

These findings support the theoretical prediction that if trade costs between cities are quite low, then different cities should have the same extent of firm selection. They also echo similar results from existing papers that larger cities do not eliminate more low-productivity manufacturing firms in France during 1994-2002, in Italy during 1995-2006, and in Japan during 1986-2014 (CDGPR, 2012; Kondo, 2017; Accetturo et al., 2018). Practically all cities in these countries were entrenched in a well-developed highway network during their study periods, which explains the absence of the selection effect, similarly to Chinese cities used in Table 6.

6.2 Results for Provinces with Underdeveloped Market Economies

Our analysis so far has focused on the five provinces with well-developed market economies. For comparison purposes, here we investigate the firm selection hypothesis for other provinces in mainland China. We choose the study period of 1998-1999 so that we can find a group of large

cities that are unconnected to a group of small cities by highway. We exclude five provinces that suffered from heavy flooding in 1998, as well as four distinct western provinces that may severely violate the assumption of the common underlying distributions of marginal cost ($G(c)$).[14]

[Table 7 here]

As displayed in Table 7, the evidence for firm selection in these provinces with underdeveloped market economies is indeed quite weak. In Panel A which pools these provinces together, $\hat{S}$ is estimated to be quite small and statistically insignificant, except that under the cutoff of 1 million it is significant but its estimated value is still lower than 1%. In Panel B, we estimate the quantile model by each individual province, using 0.5 million population as the cutoff of city size. After dropping three provinces and four municipalities that have insufficient firm observations in cities, we are left with 11 provinces for the empirical results.[15] Similarly, $\hat{S}$ is quite small in most provinces, and only in Henan province it is positive and statistically significant. Apparently, larger cities in most of these provinces do not weed out more low-productivity firms. The evidence for right-shift is still present, as $\hat{A}$ is estimated to be positive and significant in most provinces.

Even in these provinces, if choosing those sectors with the largest extent of liberalization, we might observe stronger evidence for firm selection. Indeed, there is large heterogeneity in the timing and pace of market reform across sectors in China (Qian, 2000). In order to identify the most liberalized sectors out of the 196 3-digit manufacturing sectors, we compute a ratio of the number of SOEs over the number of all above-scale firms in 1998 for each sector, and choose the 45 sectors with the ratio below 25% (the 23.5 percentile) to estimate the quantile equation (3), using the same setting as Panel A of Table 7. As displayed in Table 8, $\hat{S}$ increases to about 3% and becomes statistically significant at 5% or 1% level under different cutoffs of city size.

[Table 8 here]

[14]The five provinces that suffered from heavy flooding in 1998 are Hubei, Hunan, Jiangxi, Heilongjiang and Jilin (see the report from the United Nations at https://reliefweb.int/report/china/final-report-1998-floods-peoples-republic-china). The four western provinces, Tibet, Gansu, Qinghai and Xinjiang, are characterized by a plateau environment, very low population density and being historically underdeveloped. While occupying 42.2% of land area in China, they only accounted for 6.9% population and 2.8% GDP in 2000.

[15]The three provinces include Inner Mongolia, Hainan and Ningxia. The four municipalities include Beijing, Tianjin, Shanghai and Chongqing, which have quite small areas and each of them primarily consists of only one large city, so the segmented small cities in these municipalities are hard to find.

7 Conclusions

In this paper we revisit the firm selection hypothesis and analyze the distributions of firm productivity between small and large cities. We find that large cities eliminate a larger proportion of low-productivity firms than small cities – thus confirming the firm selection hypothesis – when the two groups of cities are not connected by controlled-access highways. When cities are connected by highways, however, larger cities do not eliminate an appreciably larger proportion of firms.

While our analysis here primarily focuses on the manufacturing sector as a whole, we also show that the results may differ across specific industries. For instance, for sectors with extraordinarily high shipping costs, city size still governs the selection effect even in the presence of well-connected highways; in regions with underdeveloped market economies, the selection effect remains weak, except for a handful of sectors as pioneers in the liberalization.

Our findings are notable for several reasons. First, we provide some of the first evidence that city size can augment the firm selection effect, consistent with the theoretical prediction of Melitz and Ottaviano (2008). Second, our results imply that the lack of firm selection found in previous studies (notably CDGPR (2012)) may result from a high level of market integration of cities in the study areas and periods. Third, our results imply that transportation costs and transportation infrastructure are an important consideration for the proper measurement of market size and identification of the spatial scope of markets.

While this paper has pointed out that transportation networks, rather than city boundaries, may define a market for the purpose of assessing the firm selection hypothesis, there are other types of effects, such as agglomeration economies and sectoral specialization, which may be more prominent at the city-level. We are examining these issues in follow-up work.

References

Accetturo, A., Di Giacinto, V., Micucci, G., & Pagnini, M. (2018). Geography, productivity, and trade: Does selection explain why some locations are more productive than others?. *Journal of Regional Science*, 58(5), 949-979.

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6), 2411-2451.

Arimoto, Y., Nakajima, K., & Okazaki, T. (2014). Sources of productivity improvement in industrial clusters: The case of the prewar Japanese silk-reeling industry. *Regional Science and Urban Economics*, 46, 27-41.

Backus, M. (2020). Why is productivity correlated with competition? *Econometrica*,88(6), 2415-2444.

Banerjee, A., Duflo, E., & Qian, N. (2020). On the road: Access to transportation infrastructure and economic growth in China. *Journal of Development Economics*, 145, 102442.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., & Zhang, Q. (2017). Roads, railroads, and decentralization of Chinese cities. *Review of Economics and Statistics*, 99(3), 435-448.

Baum-Snow, N., Henderson, J. V., Turner, M. A., Zhang, Q., & Brandt, L. (2020). Does investment in national highways help or hurt hinterland city growth? *Journal of Urban Economics*, 115, 103-124.

Behrens, K., Duranton, G., & Robert-Nicoud, F. (2014). Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy*, 122(3), 507-553.

Behrens, K., Mion, G., Murata, Y., & Suedekum, J. (2017). Spatial frictions. *Journal of Urban Economics*, 97, 40-70.

Behrens, K., & Robert-Nicoud, F. (2014). Survival of the fittest in cities: Urbanisation and inequality. *Economic Journal*, 124(581), 1371-1400.

Behrens, K., & Robert-Nicoud, F. (2015). Agglomeration theory with heterogeneous agents. *Handbook of Regional and Urban Economics*, 5, 171-245.

Chan, K. W. (2007). Misconceptions and complexities in the study of China's cities: Definitions, statistics, and implications. *Eurasian Geography and Economics*, 48(4), 383-412.

Chen, T., Gu, Y., & Zou, B. (2023). China's commuting-based metropolitan areas. Retrieved August 1, 2023 from https://ssrn.com/abstract=4052749.

Combes, P. P., Duranton, G., Gobillon, L., Puga, D., & Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6), 2543-2594.

Ding, C., & Niu, Y. (2019). Market size, competition, and firm productivity for manufacturing in China. *Regional Science and Urban Economics*, 74, 81-98.

Duranton, G. (2015). Delineating metropolitan areas: measuring spatial labour market networks through commuting patterns. In T. Watanabe, I. Uesugi and A. Ono (Ed.), The economics of interfirm networks (pp. 107-133). Tokyo: Springer Japan.

Duranton, G., & Puga, D. (2001). Nursery cities: urban diversity, process innovation, and the life cycle of products. *American Economic Review*, 91(5), 1454-1477.

Ellison, G., Glaeser, E. L., & Kerr, W. R. (2010). What causes industry agglomeration? Evidence

from coagglomeration patterns. *American Economic Review*, 100(3), 1195-1213.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. *Review of Economic Studies*, 81(3), 1046-1070.

Fujita, M., & Hu, D. (2001). Regional disparity in China 1985-1994: The effects of globalization and economic liberalization. *Annals of Regional Science*, 35(1), 3-37.

Gaubert, C. (2018). Firm sorting and agglomeration. *American Economic Review*, 108(11), 3117-53.

Gilley, B. (2001). Breaking barriers. *Far Eastern Economic Review*, 164(27), 16-16.

Gobillon, L., & Roux, S. (2010). Quantile-based inference of parametric transformations between two distributions. Processed, CREST-INSEE.

Hao, R., & Wei, Z. (2010). Fundamental causes of inland-coastal income inequality in post-reform China. *Annals of Regional Science*, 45, 181-206.

Holz, C. A. (2009). No razor's edge: Reexamining Alwyn Young's evidence for increasing inter-provincial trade barriers in China. *Review of Economics and Statistics*, 91(3), 599-616.

Jia, H., Juan, Z., Zhang, X., & Ni, A. (2004). Determination and comparison of fuel consumption for expressway post-assessment. *Journal of Jilin University*, 34(2), 298-301.

Kondo, K. (2017). Testing for agglomeration economies and firm selection in spatial productivity differences: The case of Japan. Research Institute of Economy, Trade and Industry (RIETI).

Li, S., Liu, Y., & Chen, B. (2004). Research on measures: Objects and degrees of local protection in Chinese domestic markets: an analysis based on a sample survey. Retrieved from www.hiebs.hku.hk/events_updates/pdf/lishangtong.pdf.

Li, H., & Mykhnenko, V. (2018). Urban shrinkage with Chinese characteristics. *Geographical Journal*, 184(4), 398-412.

Lu, Y., Wang, J., & Zhu, L. (2019). Place-based policies, creation, and agglomeration economies: Evidence from China's economic zone program. *American Economic Journal: Economic Policy*, 11(3), 325-360.

Melitz, M. J., & Ottaviano, G. I. (2008). Market size, trade, and productivity. *The Review of Economic Studies*, 75(1), 295-316.

Melitz, M. J., & Redding, S. J. (2014). Heterogeneous firms and trade. *Handbook of International Economics*, 4, 1-54.

Naughton, B. (2003). "How Much Can Regional Integration Do to Unify China's Markets?" in Nicholas Hope, Dennis Yang, and Mu Yang Li, eds., How Far Across the River? Chinese Policy Reform at the Millennium. Stanford: Stanford University Press, 2003. pp. 204-232.

Poncet, S. (2003). Measuring Chinese domestic and international integration. *China Economic*

Review, 14(1), 1-21.

Poncet, S. (2005). A fragmented China: Measure and determinants of Chinese domestic market disintegration. *Review of international Economics*, 13(3), 409-430.

Proost, S., & Thisse, J. F. (2019). What can be learned from spatial economics?. *Journal of Economic Literature*, 57(3), 575-643.

Qian, Y. (2000). The process of China's market transition (1978-1998): The evolutionary, historical, and comparative perspectives. *Journal of Institutional and Theoretical Economics*, 151-171.

Rupasingha, A., & Marré, A. W. (2020). Moving to the hinterlands: Agglomeration, search costs and urban to rural business migration. *Journal of Economic Geography*, 20(1), 123-153.

Syverson, C. (2004a). Market structure and productivity: A concrete example. *Journal of Political Economy*, 112(6), 1181-1222.

Syverson, C. (2004b). Product substitutability and productivity dispersion. *Review of Economics and Statistics*, 86(2), 534-550.

Wang, X., Fan, G., & Yu, J. (2009). Marketization index of China's provinces: NERI report 2008. *Social Sciences Academic Press: Beijing, China*.

Wu, S. (2005). Analysis on Decrease Factors of Bus Transportation Cost—Taking Beijing-Shijiazhuang Expressway and No. 107 State Highway as Examples. *Transportation Standardization (Jiaotong Yunshu Yanjiu)*, (7), 47.

Xing, W., & Li, S. (2011). Home bias, border effect and internal market integration in China: Evidence from inter-provincial value-added tax statistics. *Review of Development Economics*, 15(3), 491-503.

Xu, Z., & Fan, J. (2012). China's regional trade and domestic market integrations. *Review of International Economics*, 20(5), 1052-1069.

Young, A. (2000). The razor's edge: Distortions and incremental reform in the People's Republic of China. *Quarterly Journal of Economics*, 115(4), 1091-1135.

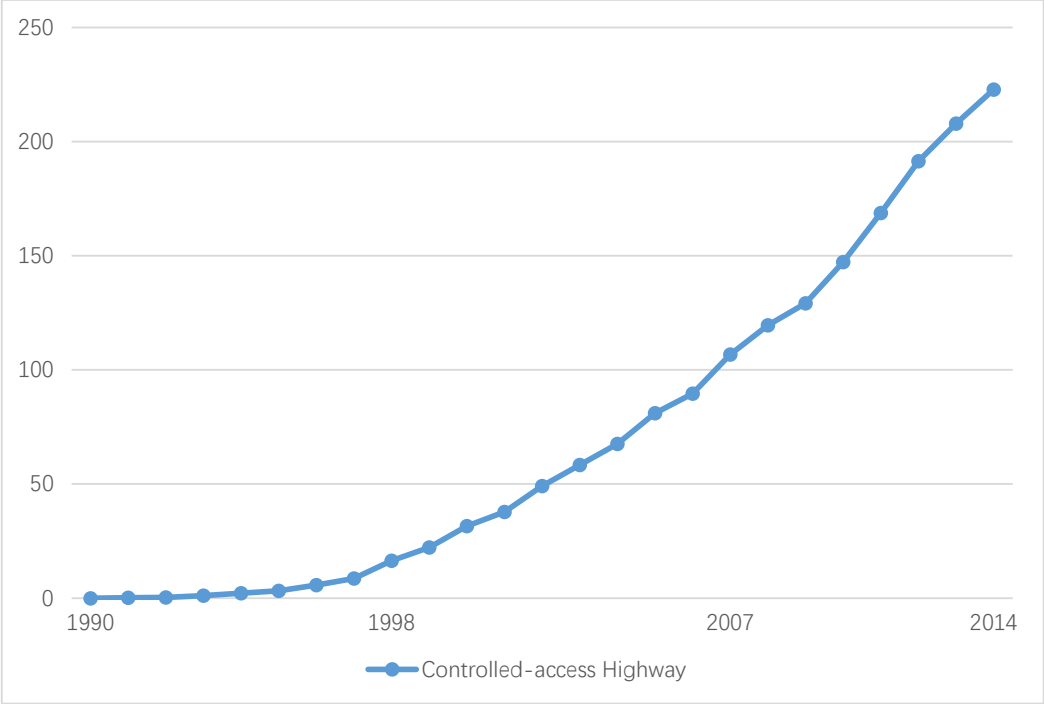

Figure 1: The Growth of Controlled-Access Highways in China (Relative to 1990).

Note: For each year, $\frac{highway_t}{highway_{1990}} - 1$ is plotted, where $highway_t$ denotes the kilometers of controlled-access highways in year *t*.

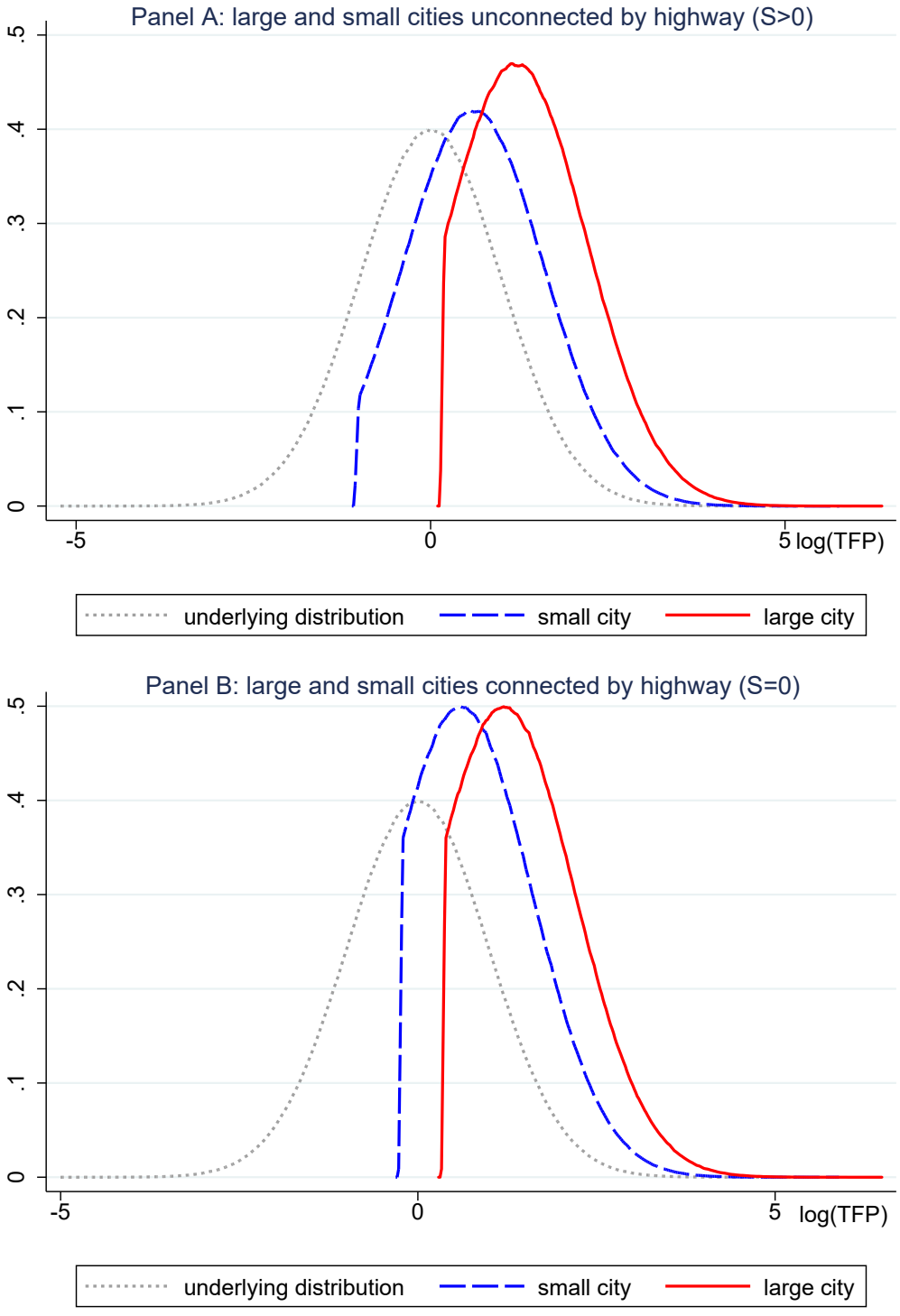

Figure 2: The Hypothetical TFP Distributions in Small and Large Cities

Note: For illustration purposes, we use the standard normal distribution as the underlying distribution. In Panel A, for the large city i and small city j, $S^i = 0.15$, $S^j = 0.05$, $A^i = 1.2$, $A^j = 0.6$, $D^i = D^j = 1$, so $S = (S^i - S^j)/(1 - S^j) = 0.105$, $D = D^i/D^j = 1$, $A = A^i - DA^j = 0.6$. In Panel B, $S^i = S^j = 0.2$, so $S = 0$; A^i, A^j, D^i and D^j are the same with Panel A, so $D = 1$ and $A = 0.6$.

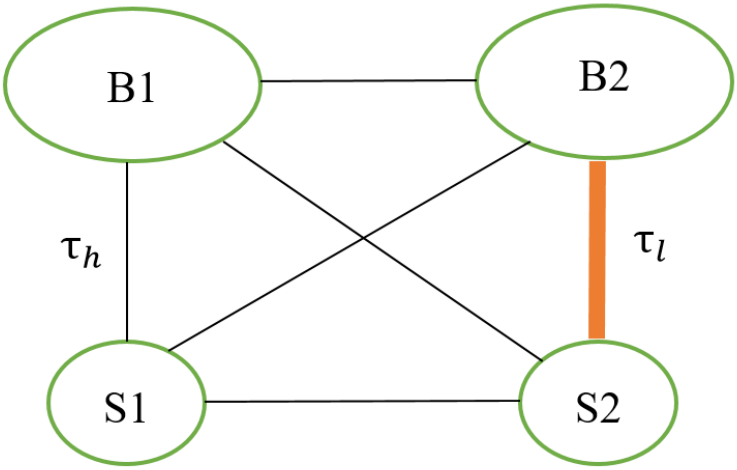

Figure 3: The Simplified Grouping of Small and Large Cities during 1998-1999

Note: (1) The large cities include B1 and B2, and the small cities include S1 and S2. (2) The thick orange line represents a controlled-access highway that corresponds to a low transportation cost (τ_l), and the thin black lines represent ordinary roads that correspond to a relatively high transportation cost (τ_h). (3) In 1999 exist many small cities like S1 that are far from highways, while in 2007 most cities are close to highways.

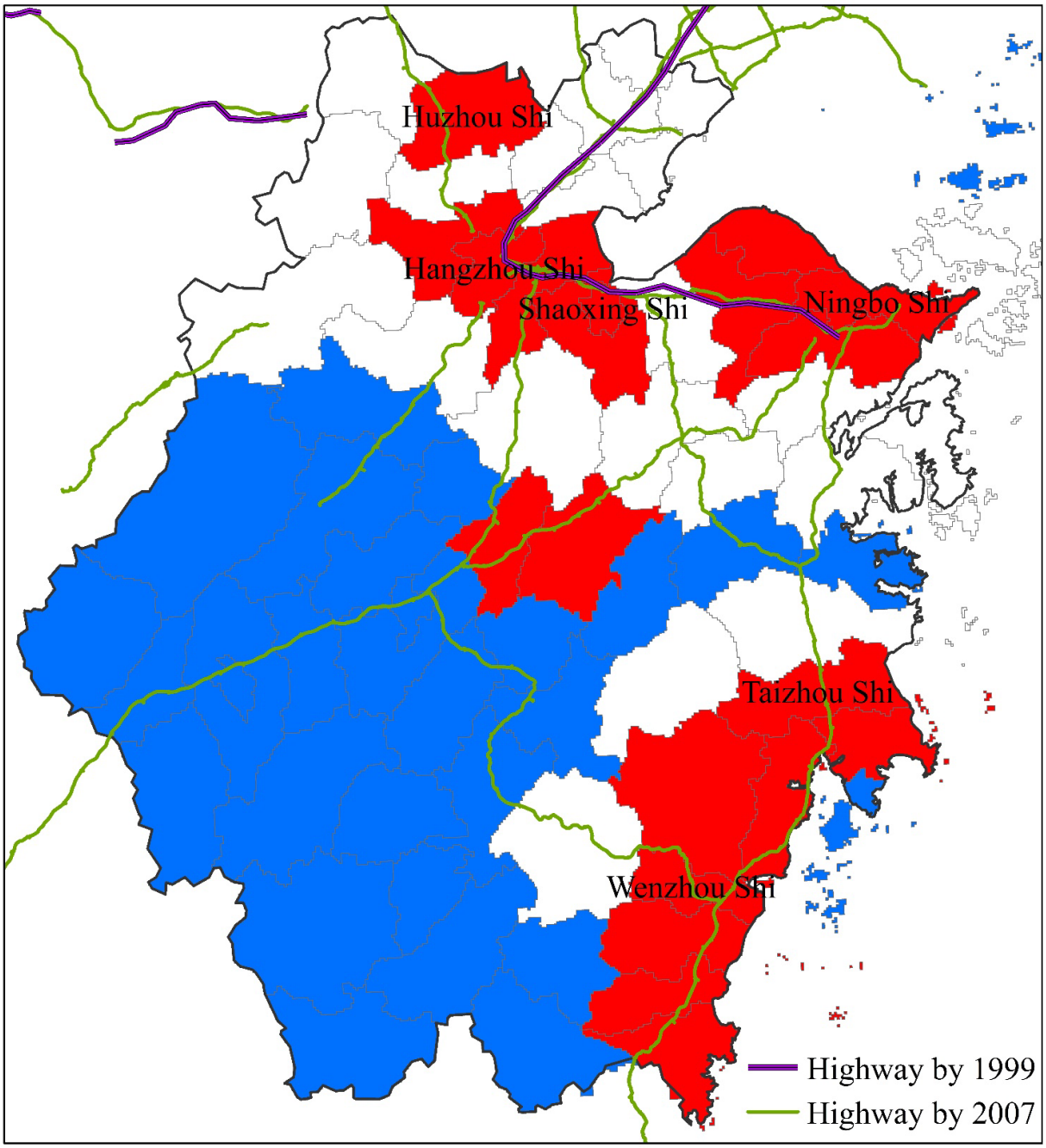

Figure 4: The Grouping of Small and Large Cities in Zhejiang Province during 1998-1999

Note: The 20 red polygons are cities larger than 500,000 population in 2000; the 35 blue polygons are cities that were smaller than 500,000, did not have any highway within 50 km from their boundaries by 1999, and did not share any boundary with any city above 1 million; the 19 white polygons are cities smaller than 500,000, but they either had highways within 50 km from their boundaries, or share boundaries with a city above 1 million. For our benchmark results in 1998-1999, we use firms in red polygons as firms in large cities, use firms in blue polygons as firms in small cities, and omit firms in white polygons. The names of the prefecture-level cities in red polygons are displayed.

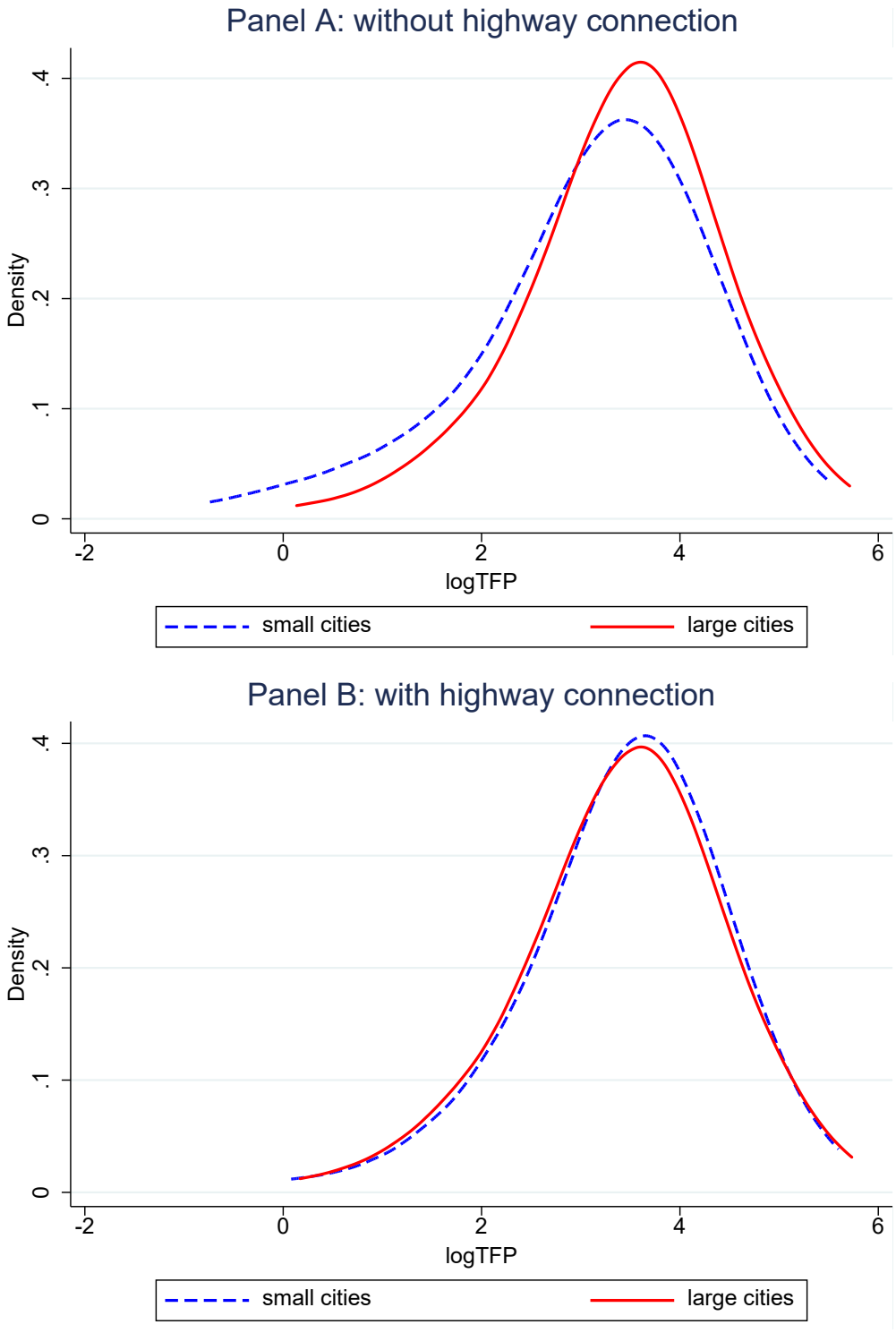

Figure 5: TFP Distributions of Large Cities and Small Cities in the Five Provinces during 1998-1999

Note: (1) The cutoff of city size is 0.5 million. (2) In Panel A, there is no highway connection between the group of large cities and the group of small cities, corresponding to the first scenario discussed in Section 4.5; in Panel B, all the cities are connected by highways, corresponding to the second scenario. A similar figure applies to each of the five provinces. (3) The bandwidth is estimated using the plug-in method.

Table 1: The Development of Market Economy by Province in 1998

Province	SOE's share of firm number		SOE's share of employment		SOE's share of value added		Marketization index		Sum of ranks
	Rank	Value	Rank	Value	Rank	Value	Rank	Value	
Zhejiang	1	15.3%	2	27.1%	1	14.8%	2	8.2	6
Jiangsu	2	18.0%	4	38.1%	3	22.8%	4	6.9	13
Fujian	3	25.9%	3	31.8%	2	22.2%	3	7.3	11
Guangdong	4	26.2%	1	24.8%	4	24.9%	1	8.2	10
Shandong	5	28.6%	5	47.0%	7	31.5%	6	6.6	23
Shanghai	6	28.9%	7	52.5%	18	45.3%	9	6.4	40
Henan	7	33.1%	8	60.0%	8	33.2%	8	6.5	31
Tianjin	8	33.9%	6	48.5%	6	30.0%	10	6.3	30
Hebei	9	42.1%	9	63.0%	10	35.4%	5	6.6	33
Anhui	10	42.5%	10	64.3%	15	42.1%	17	5.6	52
Sichuan	11	44.5%	11	68.7%	13	38.5%	18	5.6	53
Hubei	12	49.1%	13	70.4%	22	56.2%	12	6	59
Liaoning	13	50.1%	14	71.2%	12	36.7%	13	5.9	52
Chongqing	14	50.5%	19	75.5%	5	26.6%	16	5.6	54
Shanxi	15	54.3%	12	69.8%	16	42.2%	20	4.6	63
Hunan	16	56.4%	17	74.7%	9	34.6%	7	6.5	49
Beijing	17	59.5%	15	71.9%	25	59.2%	11	6.2	68
Tibet	18	59.6%	16	72.9%	28	65.1%	n.a.	n.a.	n.a.
Guangxi	19	64.1%	18	75.1%	11	35.7%	19	5.4	67
Yunnan	20	65.4%	21	78.9%	29	67.2%	27	3.7	97
Shaanxi	21	67.4%	30	86.3%	17	44.7%	22	4.4	90
Ningxia	22	69.4%	23	79.9%	27	62.2%	28	2.5	100
Jilin	23	70.2%	22	79.4%	24	58.8%	21	4.5	90
Heilongjiang	24	71.1%	24	81.1%	20	52.2%	24	4.2	92
Gansu	25	71.4%	26	84.1%	21	55.4%	23	4.3	95
Inner Mongolia	26	71.8%	29	85.7%	19	51.8%	26	3.7	100
Guizhou	27	73.7%	28	85.3%	23	57.2%	25	4.1	103
Jiangxi	28	75.0%	25	82.9%	14	39.2%	15	5.6	82
Hainan	29	75.0%	20	77.6%	26	62.0%	14	5.7	89
Xinjiang	30	79.6%	27	84.6%	31	74.8%	29	2.5	117
Qinghai	31	83.4%	31	86.4%	30	74.2%	30	1.9	122

Note: For each province, we use our ASIF data to compute the SOE's shares of firm number and employment, respectively, in the above-scale firms; when computing the SOE's share of value-added, we use our ASIF data to compute the value added of SOEs and use it as the numerator, and the denominator is the value added of the manufacturing sector from NBS (https://data.stats.gov.cn/); the marketization index comes from Wang et al. (2009). The last column, rank sum, is the sum of the four ranks from previous columns, and it suggests that the first five provinces in bold had relatively better developed market economies in 1998.

Table 2: Estimated Selection Effect in all the Five Provinces during 1998-1999, No Highway Connection between the Large and Small Cities

Cutoff of city size	$\hat{S}$ (1)	R^2 (2)	$\hat{S}$ (3)	$\hat{A}$ (4)	R^2 (5)	$\hat{S}$ (6)	$\hat{A}$ (7)	$\hat{D}$ (8)	R^2 (9)	Obs. in small cities	Obs. in big cities
1 million	0.0640*** (0.0066)	0.85	0.0390*** (0.0048)	0.1519*** (0.0252)	0.98	0.0338*** (0.0059)	0.1688*** (0.0270)	0.9779 (0.0196)	0.98	8,883	27,075
0.5 million	0.0719*** (0.0072)	0.81	0.0391*** (0.0052)	0.1940*** (0.0287)	0.98	0.0347*** (0.0065)	0.2088*** (0.0306)	0.9809 (0.0233)	0.98	6,675	39,937
75th percentile: 364,412	0.0718*** (0.0070)	0.83	0.0418*** (0.0058)	0.1823*** (0.0286)	0.97	0.0293*** (0.0069)	0.2229*** (0.0317)	0.9500** (0.0251)	0.98	6,109	43,712
Median: 202,222	0.0976*** (0.0095)	0.77	0.0460*** (0.0079)	0.2974*** (0.0355)	0.98	0.0334*** (0.0084)	0.3388*** (0.0358)	0.9503* (0.0273)	0.98	3,789	54,415

Note: (1) Small cities include all the cities that were smaller than the cutoff of city size in 2000, did not have any highway within 50 km of their administrative boundaries in the end of 1999, and did not share any boundary with any city above 1 million. Big cities include all the cities larger than any of the small cities we use. (2) in column 1-2, we only use selection (*S*) to explain the differences of TFP distributions between the big and small cities; in column 3-5, we use both selection (*S*) and agglomeration (*A*); in column 6-9, we use selection (*S*), agglomeration (*A*) and dilation (*D*). (3) In parentheses are standard errors computed from 100 bootstrapped replications. *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.

Table 3: Estimated Selection Effect in Each of the Five Provinces during 1998-1999, No Highway Connection between the Large and Small Cities

Province	$\hat{S}$ (1)	$\hat{A}$ (2)	$\widehat{D}$ (3)	R^2 (4)	Obs. in small cities	Obs. in big cities
Zhejiang	0.0220**	0.2617***	0.9978	0.99	1,289	8,592
	(0.0109)	(0.0417)	(0.0477)			
Jiangsu	0.0514**	0.0571	1.0459	0.93	1,983	10,966
	(0.0242)	(0.0803)	(0.0702)			
Guangdong	0.0554**	0.1476	0.9606	0.96	1,580	14,326
	(0.0255)	(0.0989)	(0.0640)			
Fujian	0.0298***	0.4279***	0.9345	0.99	1,248	3,521
	(0.0114)	(0.0613)	(0.0511)			
Shandong	0.0390**	0.3432***	1.0409	0.99	579	6,072
	(0.0184)	(0.0774)	(0.0625)			

Note: (1) Small cities include all the cities that were smaller than 0.5 million population in 2000, did not have any highway within 50 km of their administrative boundaries in the end of 1999, and did not share any boundary with any city above 1 million. Big cities include all the cities larger than any of the small cities we use. (2) In parentheses are standard errors computed from 100 bootstrapped replications. *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.

Table 4: Controlling for the Potential Endogeneity of Highway Construction, 1998-1999

Province	$\hat{S}$ (1)	$\hat{A}$ (2)	$\widehat{D}$ (3)	R^2 (4)	Obs. in small cities	Obs. in big cities
Five provinces	0.0280***	0.2143***	0.9245***	0.97	6,863	19,800
	(0.0075)	(0.0305)	(0.0274)			
Zhejiang	0.0192*	0.2592***	0.9384	0.98	1,289	4,578
	(0.0115)	(0.0431)	(0.0485)			
Jiangsu	0.0516**	0.0690	1.0117	0.94	1,983	6,748
	(0.0257)	(0.0834)	(0.0745)			
Guangdong	0.0459**	0.1487*	0.9028*	0.96	1,580	7,178
	(0.0235)	(0.0850)	(0.0560)			
Fujian	0.0470***	0.1787**	1.0160	0.97	1,248	1,713
	(0.0157)	(0.0728)	(0.0674)			
Shandong	0.0111	0.6033***	0.9176	0.99	579	2,879
	(0.0186)	(0.0801)	(0.0652)			

Note: (1) Small cities include all the cities that were smaller than 0.5 million population in 2000, did not have any highway within 50 km of their administrative boundaries in the end of 1999, and did not share any boundary with any city above 1 million. (2) Big cities are larger than any of the small cities we use, but they either did not have any highway by 1999, or had highways because of adjacency to the hypothetically efficient highway network. (3) In parentheses are standard errors computed from 100 bootstrapped replications. *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.

Table 5: Further Investigations of the Firm Selection Effect in the Five Provinces during 1998-1999

	$\hat{S}$ (1)	$\hat{A}$ (2)	$\hat{D}$ (3)	R^2 (4)	obs. in small cities	obs. in big cities
Using all firms instead of mono-establishment firms, small cities without highway	0.0325*** (0.0066)	0.2127*** (0.0310)	0.9726 (0.0244)	0.9773	7,010	41,532
Using the firm observations that did not export, small cities without highway	0.0209*** (0.0073)	0.2718*** (0.0321)	0.9095*** (0.0271)	0.9683	5,464	24,445
Using the 3-digit sectors whose dollar value per kilogram shipment < 2 (35th pct.), all cities have highway	0.0124* (0.0072)	-0.0062 (0.0309)	1.0364 (0.0270)	0.8256	3,640	5,462
Using the 3-digit sectors whose dollar value per kilogram shipment < 1 (23rd pct.), all cities have highway	0.0130*** (0.0046)	-0.0274 (0.0324)	1.0937*** (0.0263)	0.8930	2,365	3,363
For cities between 0.25-0.5 million, compare those with highway vs. those without any highway	0.0229*** (0.0073)	0.1227*** (0.0281)	1.0455 (0.0314)	0.9801	3,713	10,725
For cities between 0.5-0.75 million, compare those with highway vs. those without any highway	0.0232** (0.0110)	0.1390*** (0.0390)	1.1614*** (0.0486)	0.9680	1,607	2,045

Note: (1) In the first two rows, the way we group big and small cities is the same as Table 2, with the cutoff of 0.5 million. (2) In the 3rd and 4th rows, we use the cutoff of 1 million to ensure sufficient observations in small cities. (3) In parentheses are standard errors computed from 100 bootstrapped replications. *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.

Table 6: Estimated Selection Effect by Comparing Big vs. Small Cities all Connected by Highways

Province	$\hat{S}$ (1)	$\hat{A}$ (2)	$\hat{D}$ (3)	R^2 (4)	Obs. in small cities	Obs. in big cities
	Panel A: 1998-1999					
Five provinces	0.0051* (0.0027)	-0.0360* (0.0190)	1.0679* (0.0171)	0.92	7,523	26,034
Zhejiang	-0.0032 (0.0054)	-0.0210 (0.0338)	1.0328 (0.0414)	0.57	1,224	4,102
Jiangsu	0.0055 (0.0060)	0.0271 (0.0287)	1.0610 (0.0272)	0.76	2,523	6,758
Guangdong	0.0083 (0.0077)	-0.0416 (0.0352)	0.9531 (0.0410)	0.95	1,726	8,614
Fujian	-0.0015 (0.0166)	0.0964* (0.0547)	1.0698 (0.0615)	0.84	819	2,830
Shandong	0.0024 (0.0043)	-0.2050*** (0.0354)	1.0722 (0.0283)	0.98	2,293	3,216
	Panel B: 2006-2007					
Five provinces	0.0002 (0.0017)	-0.0631*** (0.0143)	1.0095 (0.0091)	0.99	57,866	100,677
Zhejiang	-0.0012 (0.0023)	0.0100 (0.0131)	1.0573*** (0.0146)	0.96	17,988	25,595
Jiangsu	-0.0019 (0.0021)	0.0591*** (0.0146)	1.0233 (0.0147)	0.98	14,893	23,248
Guangdong	0.0102* (0.0053)	0.1048 (0.0309)	1.0212 (0.0232)	0.99	3,935	35,049
Fujian	0.0068 (0.0057)	-0.0794** (0.0381)	1.0908*** (0.0276)	0.85	5,805	5,517
Shandong	-0.0039 (0.0021)	-0.0356 (0.0219)	1.0166 (0.0120)	0.88	15,249	11,271

Note: (1) In each period, we only use the cities that had highways inside their administrative boundaries at the end of the period. Among them, the cities smaller than 0.5 million are grouped as small cities, and those larger than 0.5 million are big cities. (2) In parentheses are standard errors computed from 100 bootstrapped replications. *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.

Table 7: Estimated Selection Effect in other Provinces with Underdeveloped Market Economies during 1998-1999, No Highway Connection between the Large and Small Cities

	$\hat{S}$ (1)	$\hat{A}$ (2)	$\hat{D}$ (3)	R^2 (4)	Obs. in small cities	Obs. in big cities
	Panel A: using different cutoffs of city size for all the provinces					
1 million	0.0093***	0.5074***	0.9440***	0.9842	11,348	25,998
	(0.0038)	(0.0302)	(0.0188)			
0.5 million	0.0024	0.5039***	0.9350***	0.9830	10,314	30,274
	(0.0032)	(0.0289)	(0.0177)			
75th percentile: 141635	0.0007	0.5246***	0.9296***	0.9876	6,272	44,698
	(0.0031)	(0.0313)	(0.0188)			
median: 65898	-0.0023	0.6843***	0.9063***	0.9888	3,113	54,366
	(0.0101)	(0.0490)	(0.0351)			
	Panel B: using the cutoff of 0.5 million for each individual province					
Henan	0.0162**	-0.3675***	1.1946***	0.9921	1,811	1,522
	(0.0068)	(0.0598)	(0.0483)			
Hebei	0.0114	0.7470***	0.9860	0.9906	534	1,517
	(0.0668)	(0.2573)	(0.1722)			
Anhui	0.0055	-0.1756**	1.1456**	0.9159	874	935
	(0.0152)	(0.0838)	(0.0706)			
Sichuan	0.0016	0.5894***	0.9827	0.9884	831	1,418
	(0.0089)	(0.0706)	(0.0480)			
Liaoning	0.0058	0.5722***	0.9947	0.9753	431	3,768
	(0.0476)	(0.1940)	(0.1359)			
Shanxi	-0.0083	0.0359	1.2258	0.9128	1,440	428
	(0.0311)	(0.1552)	(0.1577)			
Guangxi	0.0129	0.7019***	0.9304	0.9780	932	838
	(0.0631)	(0.2300)	(0.1377)			
Yunnan	0.0610	0.4530	1.0842	0.9371	491	535
	(0.0949)	(0.3674)	(0.2429)			
Shaanxi	-0.0152	0.8134***	0.8898	0.9887	554	921
	(0.0219)	(0.1165)	(0.0835)			
Gansu	-0.0142	0.5298***	0.9914	0.9886	831	657
	(0.0187)	(0.1007)	(0.0637)			
Guizhou	-0.0014	0.3893*	0.9580	0.9636	866	428
	(0.0666)	(0.2328)	(0.1273)			

Note: (1) Small cities include all the cities that were smaller than 0.5 million population in 2000, did not have any highway within 50 km of their administrative boundaries in the end of 1999, and did not share any boundary with any city above 1 million. Big cities include all the cities larger than any of the small cities we use. (2) In parentheses are standard errors computed from 100 bootstrapped replications. *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.

Table 8: Using Liberalized Sectors in other Provinces during 1998-1999,
No Highway Connection between the Large and Small Cities

Cutoff of city size	$\hat{S}$ (1)	$\hat{A}$ (2)	$\hat{D}$ (3)	R^2 (4)	Obs. in small cities	Obs. in big cities
1 million	0.0290***	0.1873***	1.0231	0.9618	1,301	6,860
	(0.0108)	(0.0532)	(0.0529)			
0.5 million	0.0271***	0.1504***	1.0367	0.9519	1,139	7,643
	(0.0104)	(0.0547)	(0.0532)			
75th percentile: 144,631.5	0.0391**	0.1199*	1.0588	0.9604	615	10,306
	(0.0159)	(0.0696)	(0.0658)			

Note: (1) Small cities include all the cities that were smaller than 0.5 million population in 2000, did not have any highway within 50 km of their administrative boundaries in the end of 1999, and did not share any boundary with any city above 1 million. Big cities include all the cities larger than any of the small cities we use. (2) We exclude five provinces that were heavily flooded in 1998, including Hubei, Hunan, Jiangxi, Heilongjiang and Jilin. We also exclude four outlier provinces in Western China, including Tibet, Gansu, Qinghai, and Xinjiang. (3) In parentheses are standard errors computed from 100 bootstrapped replications.

Online Appendix

A Proof for the Firm Selection Hypothesis

Consider any two cities i and j in province u such that $L_u^i > L_u^j$. The free entry condition (2) for the two cities are:

$$\frac{L_u^i}{4\gamma}\int_0^{\overline{p}_u^i}\left(\overline{p}_u^i - c\right)^2 g\left(c\right)dc + \frac{L_u^j}{4\gamma}\int_0^{\overline{p}_u^j/\tau_1}\left(\overline{p}_u^j - \tau_1 c\right)^2 g\left(c\right)dc + \sum_{h\neq i, h\neq j}\frac{L_u^h}{4\gamma}\int_0^{\overline{p}_u^h/\tau_1}\left(\overline{p}_u^h - \tau_1 c\right)^2 g\left(c\right)dc + \sum_{v\neq u}\sum_k \frac{L_v^k}{4\gamma}\int_0^{\overline{p}_v^k/\tau_2}\left(\overline{p}_v^k - \tau_2 c\right)^2 g(c)dc = f_E, \tag{A1}$$

$$\frac{L_u^j}{4\gamma}\int_0^{\overline{p}_u^j}\left(\overline{p}_u^j - c\right)^2 g\left(c\right)dc + \frac{L_u^i}{4\gamma}\int_0^{\overline{p}_u^i/\tau_1}\left(\overline{p}_u^i - \tau_1 c\right)^2 g\left(c\right)dc + \sum_{h\neq i, h\neq j}\frac{L_u^h}{4\gamma}\int_0^{\overline{p}_u^h/\tau_1}\left(\overline{p}_u^h - \tau_1 c\right)^2 g\left(c\right)dc + \sum_{v\neq u}\sum_k \frac{L_v^k}{4\gamma}\int_0^{\overline{p}_v^k/\tau_2}\left(\overline{p}_v^k - \tau_2 c\right)^2 g(c)dc = f_E. \tag{A2}$$

Subtract (A2) from (A1) and we have

$$L_u^i f\left(\overline{p}_u^i, \tau_1\right) = L_u^j f\left(\overline{p}_u^j, \tau_1\right), \tag{A3}$$

where $f\left(x, \tau_1\right) = \int_0^x \left(x - c\right)^2 g\left(c\right)dc - \int_0^{x/\tau_1}\left(x - \tau_1 c\right)^2 g\left(c\right)dc$.

If $\tau_1 > 1$, we have $\partial f\left(x, \tau_1\right)/\partial x > 0$, and since $L_u^i > L_u^j$, we must have $\overline{p}_u^i < \overline{p}_u^j$. Given $S_u^i = 1 - G(\overline{p}_u^i)$ and $S_u^j = 1 - G(\overline{p}_u^j)$, we obtain $S_u^i > S_u^j$.

If $\tau_1 = 1$, then $S_u^i = S_u^j$, as shown in Appendix B of CDGPR (2012). It suggests that cities in the same province share a common threshold of productivity, so they do not have any differences in terms of selection effects.

B Description and Checking of the Firm-Level Data

The firm-level data we use comes from China's Annual Survey of Industrial Firms (ASIF) during 1998-2007, collected by China's National Bureau of Statistics (NBS). Included in the data are the so-called above-scale firms, which consist of all state-owned firms enterprises (SOEs) and those non-state-owned firms with sales exceeding 5 million RMB in a year, in the sectors of mining, manufacturing and public utilities. The data reported by the above-scale firms are rather reliable since they have independent accounting systems and are subject to a regular reporting system by

NBS (Holz, 2008). Along with China's rapidly growing economy, the industrial sectors, which make up a large component of the Chinese economy[1], had expanded substantially, with the number of above-scale firms increasing from 165,118 in 1998 to 336,768 in 2007; employment from 56.44 to 78.75 million; value added from 1.94 to 11.70 trillion RMB; and sales from 6.41 to 39.97 trillion RMB.

Comparing the ASIF data to the first National Economic Census in 2004 that surveyed all registered industrial firms, we find that, the above-scale firms employed approximately 78% of workers, produced 90% of the national industrial output, and made up 20% of all firms.

In manufacturing sectors that this paper investigates, above-scale firms also made up 20% of all firms in 2004, but this share should be higher in 1998, because the number of SOEs, all of which are included in the ASIF data, shrank by 47% during 1998-2004 under the market reform. Assuming the number of non-state-owned above-scale firms increased by the same proportion as the number of below-scale firms during 1998-2004, our estimated share of the above-scale firms in all manufacturing firms is approximately 27% in 1998.[2]

The noise in the firm-level data could generate notable attenuation bias of the estimated firm selection effect in the quantile approach (Ding and Niu, 2019), so we made the following efforts to reduce the noise.

B.1 Checking the Raw Data

While a growing number of studies use the ASIF data, their summary statistics are not always the same, perhaps because they obtain the data from different vendors. We obtained the ASIF data from two data vendors separately, hereinafter referred to as version 1 and version 2. Using the firm identifier (the unified code for legal person) assigned by China's NBS, as well as other information including firm name, legal representative, etc., we merged the two data versions and checked whether the same firm reports the same values in the two versions. The main discrepancies pertain to a small number of firms' county-level codes (the first 6 digits of the administrative-division code, which represent a county, county-level city or city district), and the

[1] In 2000, for example, the industrial sector contributed 40.4% of Chinese GDP, while the construction sector contributed only 5.6% of GDP, and the primary and tertiary sectors produced 15.1% and 39.0% of GDP, respectively (data source: NBS, 2010, *China Compendium of Statistics 1949-2008*, China Statistical Press).

[2] In 2004, the number of all manufacturing firms is 1,258,586, consisting of 256,999 above-scale firms (27,071 SOEs + 229,928 non-state-owned firms) and 1,001,587 below-scale firms. In 1998, the number of above-scale firms in manufacturing sectors is 149,674 (57,139 SOEs + 92,535 non-state-owned firms), while the number of below-scale firms is unknown. Assuming the number of below-scale firms grow by the same proportion as the number of non-state-owned firms during 1998-2004 (229,928 / 92,535 = 2.4848), we estimate that in 1998 there are 403,091 below-scale firms (1,001,587 / 2.4848 = 403,091), and the share of above-scale firms in all manufacturing firms is approximately 27.1% (149,674 / (149,674 + 394,598) = 0.2708).

remaining variables have almost exactly the same values. The details are as follows.

The number of firms that have different county-level codes across the two versions is 593 out of 165118 in 1998, 464 out of 162033 in 1999, 461 out of 162887 in 2000, 34 out of 181557 in 2002, 46 out of 196222 in 2003, 3 out of 301961 in 2006, and 5 out of 336768 in 2007. Checking the firm addresses, we find that for these years, the county-level codes in version 1 are always correct, so these codes are used.

The exception is the year of 2001, in which the county-level codes in version 1 are all wrong, so we use the codes from version 2. As version 2 misses 2225 firm observations in this year, we use the zip codes of these firms to determine their county-level codes.[3]

For the year of 2004, version 1 has 2618 more observations than version 2. We found that some firms appear multiple times in version 1, and the number of firms in version 2 is exactly the same as the number reported in China Statistical Yearbook, so we use version 2 for 2004. Neither version reports value added in this year, and we computed it as: *value added = output - intermediate input + value added tax payable*.

For the year of 2005, version 1 has 1792 fewer observations than version 2. The number of firms in version 2 is exactly the same as the number in China Statistical Yearbook, so we use version 2 for 2005.

Then we compare the statistics of our firm-level data with those published by NBS as well as the firm-level data used in Table A1 of Brandt et al. (2014). Our Table B1 indicates that our data are equal or very close to the NBS data, and in a few cases such as 2001, our data are closer to the NBS data than Brandt et al. (2014).

B.2 Harmonizing Industry Classification

We use the information on industry classification in the estimation of the production function and the investigation of the firm selection effect by sectors. The four-digit Chinese Industry Classification (CIC) was revised since 2003 in our study period, and a consistent classification over the years was constructed by Brandt et al. (2014), i.e., the original industry codes were adjusted and became consistent across the entire period.[4] However, we found that the harmonized classification overlooked five 4-digit industry codes began with "171" (processing of fibrous materials)

[3] We found that the county-level code has fewer errors than the zip code, perhaps because firms reported their county-level codes more carefully than zip codes, or the government inspected county-level codes more carefully after receiving the data from firms, so we always use the county-level code to identify each firm's location whenever possible.

[4] We downloaded all the codes used in Brandt et al. (2014 and 2019) at https://feb.kuleuven.be/public/N07057/CHINA/appendix/.

during 1998-2002, which contain 5027 firm observations. After checking the definitions of the five industries and tracking how the firms reported their industry codes before and after the CIC adjustment in 2003, we added the five industries to the harmonized classification of Brandt et al. (2014) by adjusting the original industry code 1712 to 1721, adjusting 1713 and 1714 to 1730, adjusting 1719 to 1711, and leaving the original code 1711 unchanged. This updated industry classification is used in this research.

B.3 Checking the County-Level Code

We primarily rely on the county-level code (the first six digits of the administrative division code) to identify each firm's location.[5] Comparing to the county-level code published by China's Ministry of Civil Affairs (MCA),[6] we found many errors of this code in our data and made the corrections on most of them.

One type of the errors is that the last two digits of some county-level code were falsely reported as "00".[7] The number of firm observations having this problem is 5021 (from 193 county-level divisions), including 906 firms (from 33 county-level divisions) in 1998, 135 (from 3 divisions) in 1999, 435 (from 12 divisions) in 2000, 1382 (from 91 divisions) in 2001, 1956 (from 36 divisions) in 2002, 28 (from 1 division) in 2003, 62 (from 5 divisions) in 2005, 26 (from 3 divisions) in 2006, and 91 (from 9 divisions) in 2007.

Besides the above firm observations, many other observations still contain invalid county-level codes that could not be found from the codes published by MCA.[8] The number of these firm observations is 17872 (from 107 county-level divisions), including 10683 firms (from 40 divisions) in 1998, 1334 (from 27 divisions) in 1999, 316 (from 18 divisions) in 2000, 213 (from 8 divisions) in 2001, 51 (from 1 division) in 2002, 180 (from 7 divisions) in 2003, 37 (2 divisions) in 2006, and 5058 (4 divisions) in 2007.

Regarding the above types of errors, we corrected the county-level codes according to the first four digits of the zip codes, each of which usually corresponds to a unique county-level

[5] While since 2004 the division code in our data contains 12 digits which correspond to a unique neighborhood committee (*juweihui*) or administrative village, during 1998-2003 the division code only reports six digits. The 1st and 2nd digits represent the province-level administrative division, the 3rd and 4th digits represent the prefecture-level division, and the 5th and 6th digits represent the county-level division.

[6] We use the county-level codes published by MCA in various years at https://www.mca.gov.cn/article/sj/xzqh/1980/.

[7] During our study period, only in the city proper of Dongguan Shi and Zhongshan Shi should the last two digits of the county-level code be "00", and reporting "00" in the last two digits for any other county-level divisions is false.

[8] One of the reasons for such differences is that some firms reported outdated county-level codes, which could change due to the adjustment of administrative divisions.

division,[9] as well as other information including the names of prefecture and county.

B.4 Matching Firms over Years

Estimating the production function and computing the firm's average TFP over years require matching firms over time and building a panel dataset in advance. To do so, we use and improve the approach of Brandt et al. (2014).

We match firms in two consecutive years as follows. First, we match firms using the firm identifier assigned by NBS. For the unmatched firms, we use the firm name to find a match. These two steps are the same with Brandt et al. (2014).

In the third step, we use the combination of the county-level code, the name of the legal representative and the 2-digit industry code to link the unmatched firms in previous steps. This is different from Brandt et al. (2014), who use the combination of the name of legal representative and the prefecture-level code (the first four digits of the administrative division code) to link the unmatched firms in previous steps, and their problem is that the same legal person sometimes own multiple firms in the same prefecture, leading to different firms being identified as the same one.

The third step of the matching could be influenced by the fact that, for firms that did not relocate, their county-level codes might change due to administrative division adjustments, including county changed to county-level city, county to city district, prefecture (*diqu*) to prefecture *shi*, etc. Therefore, based on the county-level codes published by MCA in various years, we created harmonized county-level codes that are consistent across the entire period, and used them in the third step of the matching. However, the changes of the county-level codes below the county-level still could not be addressed. For example, a town belonging to County A in the first year became part of County B in the second year. Therefore, in the fourth step, we use the combination of the first four digits of zip code, as well as the name of the legal representative and the 2-digit industry code, to link unmatched firms from the previous steps.

In the fifth step, we use the combination of the county-level code, the 2-digit industry code and phone number to link the unmatched firms in previous steps. This is different from Brandt et al. (2014), who use the combination of the prefecture-level code, the 3-digit industry code and phone number to link the unmatched firms, and their problem is that some firms changed their

[9]The location of the zip codes and the zip codes of the county-level divisions can be found at www.youbianku.com. Besides searching on this website, we also used the firm observations with valid county-level codes to build a correspondence table between county-level codes and zip codes, and used the correspondence table to find the correct county-level codes according to the zip codes.

3-digit industry code back and forth. We use the 2-digit industry code instead, which was rarely changed by firms.

Brandt et al. (2014) has one more step to link the unmatched firms in previous steps. The identifier is the combination of the founding year, county-level code, 4-digit industry code, name of town, and the first of the three main products. We find that this step sometimes recognizes different firms as the same one, so we do not use it.

As some firms might disappear from the sample and re-enter later, similar to Brandt et al. (2014), we subsequently match remaining observations in data files two years apart (i.e., matching year t with year $t+2$). We added an additional step: for those firms that still appear only once in the data, we also use all other years to find matches (i.e., matching year t with $t+3$, $t+4$, ...).

B.5 Estimating the Real Values of Capital Stock, Output and Input

We need to use the firms' capital stock in the estimation of production function. As the firms only reported the nominal values of their capital stock, we follow Brandt et al. (2014)'s approach to estimate their real capital stock.

We also use the approach of Brandt et al. (2014) to deflate output and intermediate input. This approach uses the nominal and real values of output reported by each firm to construct an output deflator at the 4-digit industry level during 1998-2003. After 2003, firms did not report the real values of output anymore, and the 2-digit ex-factory price index from the China Statistical Yearbook is used as the output deflator during 2004-2007. With this output deflator and the input share calculated from the 2002 National Input-Output table, an input deflator was built following Brandt et al. (2019). Finally, we compute the real value added as: *real value added = real output - real intermediate input + real value added tax payable*.

B.6 Identifying the State-Owned Enterprises

We need to use an indicator for State-Owned Enterprises (SOEs) when estimating the production function and investigating the firm selection effect. Two main methods have been used to identify the SOEs in the ASIF data. The first method uses the information on the registered capital, as each firm reported its paid-up capital for six types of owners: state, collective, individual, legal person, Hong Kong-Macau-Taiwan, and foreign. For instance, the firms with positive capital from the state are identified as SOEs in Yu (2015), and the firms with more than 30% of capital from the state are identified as SOEs in Huang et al. (2017). The problem of this method,

however, is the difficulty of tracing the sources of the capital from the legal person, which can capture a wide range of possibilities including state-controlled shareholding companies and private subsidiaries (Brandt et al., 2014). The share of legal person in all paid-up capital increased from 18% in 1998 to 33% in 2007.

The second method identifies SOEs according to the 23 types of the firm's legal registration. This method is used by, for example, Brandt et al. (2012). While the types of state-owned enterprise (registration code 110), state-owned jointly operated enterprise (141) and state-owned LLC (151) are usually recognized as SOEs, it is difficult to identify SOEs from some other types such as joint-stock cooperative enterprise (130), state and collectively owned jointly operated enterprise (143), stock limited company (160), etc. Moreover, many SOEs are legally registered as foreign firms, limited-liability firms, or publicly traded firms (Hsieh and Song, 2015). Dollar and Wei (2007) also suggest that some former SOEs do not change their registered ownership type even after ownership restructuring.

In our data, an indicator of the shareholding status suggests two types of state control, namely, absolute control (the state share exceeding 50%) and relative control (the state share less than 50% but being the largest shareholder). In 1998-2005, the indicator's value of 1 denotes absolute control, and 2 denotes relative control; in 2006-2007, the value of 1 denotes both absolute control and relative control. Therefore, we identify the firms of the absolute control and the relative control as SOEs. This definition is probably the same with NBS, as Table B2 shows that the number of SOEs from our data is almost the same with the NBS data, except in 1998 when our number is 5% higher than the NBS's number.

According to our definition, 72.7% of SOEs observations are registered as state-owned enterprise (110), 10.2% are other LLC (159), 4.4% stock limited company (160), 3.0% state-owned LLC (151), 2.8% Chinese-foreign jointly owned enterprise (310), 2.8% Chinese-HMT jointly owned enterprise (210), and the rest types are all below 1%. This is consistent with Hsieh and Song (2015)'s finding that SOEs could be registered as foreign firms, limited-liability firms, or publicly traded firms. In the 389,905 (94.2%) SOE observations that reported positive values of paid-up capital, 76.1% reported that the state share was greater than 50%, and 76.7% reported that the state hold the largest share. While how much capital labeled as the legal person is actually from the state remains unknown, among SOE observations with positive paid-up capital, the capital share of the state and the legal person is larger than 50% in 93.4% observations, and the share of the state and the legal person is larger than any other type in 93.6% of observations.

C Estimation of the Production Function

Ordinary least squares estimation of equation (4) may exhibit bias due to the endogeneity of inputs – for instance, a firm may adjust its input use after observing its productivity shocks. To handle this problem, Olley and Pakes (1996) model firm investment as a function of productivity and capital, and thereby use investment as a proxy for unobservable productivity shocks. However, this is problematic because investment is typically lumpy and it contains many "zero" observations. To solve this problem, Levinsohn and Petrin (2003) model intermediate inputs, instead of investment, as a function of productivity and capital, since intermediate inputs may be more responsive to productivity shocks than investment. However, these two estimators will fail to identify the labor coefficient if labor input and the proxy function for the unobserved productivity are perfectly collinear. Ackerberg et al. (2015) propose a correction based on inverting the intermediate demand functions conditional on labor inputs. We add an SOE dummy into the demand function for intermediate inputs, as SOEs might have different decision behaviors than non-SOEs. Then we estimate function (4) separately by each of the 28 2-digit sectors. The estimation results of the production function for each sector are displayed in Table C1.

D Description and Checking of the Highway Data

We first downloaded the GIS (geographic information system) data of China's controlled access highways in 1999, 2005 and 2010 that were used in Baum-Snow et al. (2017) and Baum-Snow et al. (2020).[10] These data were created by digitizing the published maps of China. After carefully checking the data, we found and corrected the following errors.

First, the highways under construction are falsely used as completed highways. The published maps of China use dashed lines to represent highways under construction, and use continuous lines to represent highways that have been completed and open to traffic. However, both types of lines were used as completed highways in the GIS data we downloaded. In particular, the data of 1999 has a significant part of highways under construction. We deleted those highways under construction in the GIS data.

Second, the published map of China in 1999 contains many errors of the highways. Some highway segments are missed on the map, some highway segments completed on the map were actually under construction, and a few highway segments on the map did not exist in reality at

[10]To download the data, we visited the website address of https://matthewturner.org/research.htm, found the download link, and downloaded the data from https://drive.google.com/file/d/156qjjt2R9ADiVzqNjdJ9z5hT_AkmoSAb/view?usp=sharing.

all. These errors were found in the following ways. We collected the Yearbook of China Transportation & Communications (YCTC, *zhongguo jiaotong nianjian*) in various years, which reports the highway segments completed in each year, and each segment's start point, end point and length. Comparing highways on YCTC during 1990-1999 to the GIS data of highways in 1999 helped to find errors in the data. To double-check the possible errors, we searched online for the news on the completion of corresponding highway segments, and checked the corresponding provincial gazetteers when needed. Once confirming the errors in the GIS data, we made corrections on the GIS map based on the information from YCTC, the news and provincial gazetteers.

In addition, the GIS data for 1999 we downloaded displayed highways by January of that year, and we revised the data so that it displayed the highways by the end of 1999.

The GIS data after our corrections and revisions indicate that the total length of highways at the end of 1999 is 11564 km, very close to 11605 km reported by NBS. For comparison, the original GIS data we downloaded indicate that the total length of highways by January 1999 is 13983 km, while the NBS reported only 6258 km of highways by the end of 1998. The length of highways that exist in our GIS data but do not exist in the original GIS data is 3284 km, the length of highways that exist in the original GIS data but do not exist in our GIS data is 5703 km, and the length of highways that exist in both data is 8280 km.

In addition to the above highway data, from a data vendor, we also obtained another GIS data of China's controlled access highways at the end of 2007 from Baidu Map. After checking the data, we found it quite accurate.

E Measuring City Size in China

There is no well-recognized ready-made dataset of city size in China. While the population data of administratively designated *shi* (also known as administratively designated city, including prefecture-level city and county-level city) published in China Urban Statistical Yearbook (CUSY) and China Urban Construction Statistical Yearbook (CUCSY) in various years are sometimes used to measure city size, they have several limitations.

In terms of coverage, the yearbooks omit cities located in administratively designated *xian* (also known as administratively designated county and county equivalent), which contain 23.1% of China's urban population in 2000, for instance.

Moreover, the population indicators in the two yearbooks have several problems. First, most of the population indicators, including total population and non-agricultural population in CUSY and urban population in CUCSY, are constructed based on *hukou* (registered residence), which

has increasingly large difference from the actual residence. In 2000, for instance, 10.2% of the Chinese population live in townships that differ from their *hukou*; as a large part of such population move from rural areas and small cities to large cities, using *hukou* population would generate bias in the measure of city size in different directions and magnitudes for different cities. Second, *shi* contains much rural population, so using the total population in CUSY would overestimate city size to different extents for different cities. In 2000, the rural population accounts for 42.7% of the total population in all the 653 *shi*. Third, the population data in CUCSY contain much noise as they exhibit implausibly large fluctuations over years for many *shi*.

In this paper, we use the urban population in each *shi* or *xian* (referred to as city in this paper) as the measure of population size of each city, as discussed in Section 4.4. Using the urban area of each *shi* or *xian* as the spatial scope of the city is supported by several empirical studies which use mobile phone data or Baidu Map's commuting information in the 2010s and report that most commuting in China did not cross the boundaries of *shi* and *xian* (Ding et al., 2015; Wang et al., 2018; Zhao, et al., 2019; Chen et al., 2023). In 1998-1999, our main study period, the commuting distance should be even shorter.

We also found several cases where the major urban areas in several adjacent cities expanded across their administrative boundaries and became contiguous, and their labor markets are possibly integrated. We used two data of land use and night light. The land-use data is from the Chinese Land Use Cover Change 100-Meters Grid Dataset, obtained from the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (www.resdc.cn). It identifies 25 land-use types based on the Landsat TM/ETM satellite image, and we use both types of "urban land" (#51) and "other construction land" (#53) as urban area in this research.[11] The night-light data is from the website of the National Oceanic Atmospheric Administration. Similar to Dingel et al. (2021), we use the grids above the 30 light-intensity as urban area. If both data of land use and night light show that the two adjacent cities' major urban areas are contiguous, then we list them in Table E1. Most of these pairs of cities involve a very large city and a quite small city. When we divide cities into a group of large cities and a group of small cities as discussed in Section 4.5, if one city in Table E1 is classified as a large city, then the other city in the city pair is also treated as a large city even if its own city size is below the cutoff of city size.

We require that a city should have no less than 10,000 population, and thereby identify 2,208 cities in 2000. Their average size is 207.5 thousand, and the largest city (Shanghai) reaches 13.5

[11] Many development zones are classified as "other construction land."

million, as reported in Table E2. The five provinces with well-developed market economies contain 426 cities, with the average size of 350.4 thousand. In each of the five provinces, the number of cities ranges between 74 and 109, and the average size ranges between 210.4 thousand and 484.0 thousand.

F Measuring the Shipping Costs for 3-digit Manufacturing Sectors in China

We do not find any data on the shipping costs for manufacturing sectors in China around 2000. Similar to Syverson (2004) that uses value-to-weight ratio to measure shipping costs, we use the ratio in the United States to proxy the shipping costs in China.

First, we downloaded Table 13 from the 2002 Commodity Flow Survey (CFS) of the United States.[12] It reports the value, tons, ton-miles and the average miles per shipment for the goods shipped within the United States at the 4-digit level of the Standard Classification of Transported Goods (SCTG) codes.

Second, we managed to find China's closest 3-digit codes of manufacturing sectors for each of the United States' 4-digit commodity codes. To do so, we compared the text description of manufacturing sectors in China's industrial classification with the description of the SCTG codes in the United States.[13] Finally, out of the 258 4-digit commodities (0411-4099) in manufacturing sectors, we found a unique 3-digit industrial code for each of the 184 4-digit commodities, we found two to seven 3-digit industrial codes for each of the 65 4-digit commodities, and we failed to find any suitable industrial sectors for nine 4-digit commodities. Among China's 172 3-digit manufacturing sectors during 1998-1999, 148 sectors found at least one corresponding 4-digit commodity in the United States; the 24 3-digit sectors that did not find any suitable 4-digit commodity include the repair of certain machines and equipment (3-digit code: 298, 358, 368, 378, 408, 418 and 428), the category of "other products" in some 2-digit sectors (159, 249, 379, and 429), some Chinese specialties such as Chinese Herbal Medicines and Proprietary Chinese Medicines (273) and Bamboo, Rattan, Palm and Grass Products (204), etc.

Third, for each of China's 3-digit manufacturing sectors, we compute a weighted average of

[12] We downloaded Table 13 at https://www.bts.gov/archive/publications/commodity_flow_survey/2002/united_states/index. We did not use the 1997 Commodity Flow Survey as the level at which it reports shipment costs is the 3-digit commodities, while the 2002 CFS reports the 4-digit commodities, whose finer classification helps to measure shipping costs across sectors more accurately (the number of 3-digit commodities in manufacturing sectors is 118, while the number of 4-digit commodities is 258).

[13] We found the description of China's industrial classification at http://www.jincao.com/fa/09/law09.27.htm and found the description of SCTG codes at https://bhs.econ.census.gov/bhsphpext/brdsearch/scs_code.html.

value per ton of the corresponding 4-digit commodities. The tons of the 4-digit commodities is used as the weight. After dropping a few sectors whose corresponding 4-digit commodities do not report tons or values, we find quite large dispersion of value-per-ton in the remaining 144 3-digit sectors: the mean is 19.4 while the standard deviation reaches 101.4; the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles are 0.27, 1.22, 3.89, 10.08 and 19.74, respectively.

After inspecting the above data, we found that the variable of value-per-ton indeed serves as a good proxy of shipping costs. For instance, the sectors of construction materials, which generally have quite high shipping costs, obtain very low values of value-per-ton: Masonry, Lime and Light Building Materials (313), Cement (311), and Products of Cement and Asbestos Cement (312) report the three lowest values of 0.01, 0.07 and 0.08, respectively; Apparel Manufacturing (181), which have relatively low shipping costs, reports a high value of 19.7.

Meanwhile, we also note some factors other than value-per-ton influence shipping costs. For instance, shipping time may be crucial to some sectors like ready-mixed concrete and foods with a short shelf-life, and some pharmaceutical preparations require cryogenic transportation. Therefore, while sectors with high values of value-per-ton may not necessarily have low shipping costs, we are more confident that sectors with low values of value-per-ton should have high shipping costs.

G Investigating the Firm Selection Effect by Subsector

We investigate the selection effect by five subsectors, where each subsector consists of one 2-digit sector or a few 2-digit sectors that produce similar products. Specifically, the first subsector consists of Textile (2-digit SIC: 17), Garments & Other Fiber Products (18), and Leather, Furs, Down & Related Products (19); the second subsector consists of Timber Processing, Bamboo, Cane, Palm Fiber & Straw Products (20), and Furniture Manufacturing (21); the third and fourth subsectors are Raw Chemical Materials & Chemical Products (26) and Nonmetal Mineral Products (31), respectively; the fifth subsector consists of General Equipment Manufacturing (35), Special Equipment Manufacturing (36), and Communications Equipment, Computers and other Electronic Equipment Manufacturing (40).

As displayed in Table G1, we still find evidence for firm selection in subsectors. In four subsectors, we obtain positive values of $\hat{S}$, ranging 1.8% to 4.9%. While with the small sample size, particularly in the group of small cities, the standard errors of $\hat{S}$ are much larger than those in Table 2, $\hat{S}$ is still significant at the 5% level for three subsectors.

References

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6), 2411-2451.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., & Zhang, Q. (2017). Roads, railroads, and decentralization of Chinese cities. *Review of Economics and Statistics*, 99(3), 435-448.

Baum-Snow, N., Henderson, J. V., Turner, M. A., Zhang, Q., & Brandt, L. (2020). Does investment in national highways help or hurt hinterland city growth? *Journal of Urban Economics*, 115, 103124.

Brandt, L., Van Biesebroeck, J., Wang, L., & Zhang, Y. (2019). WTO accession and performance of Chinese manufacturing firms: Corrigendum. *American Economic Review*, 109(4), 1616-21.

Brandt, L., Van Biesebroeck, J., & Zhang, Y. (2012). Creative accounting or creative destruction? Firm-level productivity growth in Chinese manufacturing. *Journal of Development Economics*, 97(2), 339-351.

Brandt, L., Van Biesebroeck, J., & Zhang, Y. (2014). Challenges of working with the Chinese NBS firm-level data. *China Economic Review*, 30, 339-352.

Chen, T., Gu, Y., & Zou, B. (2023). China's commuting-based metropolitan areas. Retrieved August 1, 2023, from https://ssrn.com/abstract=4052749.

Combes, P. P., Duranton, G., Gobillon, L., Puga, D., & Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80(6), 2543-2594.

Ding, C., & Niu, Y. (2019). Market size, competition, and firm productivity for manufacturing in China. *Regional Science and Urban Economics*, 74, 81-98.

Ding, L., Niu, X., & Song, X. (2015). Liyong shouji shuju shibie Shanghai zhongxincheng de tongqinqu [Identifying the commuting area of Shanghai central city using mobile phone data], *Chengshi guihua*, 9,100-106.

Dingel, J. I., Miscio, A., & Davis, D. R. (2021). Cities, lights, and skills in developing economies. *Journal of Urban Economics*, 125, 103174.

Dollar, D., & Wei, S. J. (2007). Das (Wasted) Kapital: Firm Ownership and Investment Efficiency in China. *IMF Working Paper No. 2007/009*.

Holz, C. (2008). How Can a Subset of Industry Produce More Output than All of Industry? Retrieved from http://carstenholz.people.ust.hk/CarstenHolz-industry-stats-07-web-27Nov08.pdf.

Hsieh, C. T., & Song, Z. M. (2015). Grasp the Large, Let Go of the Small: The Transformation of the State Sector in China. *Brookings Papers on Economic Activity*, Spring, 295-346.

Huang, Z., Li, L., Ma, G., & Xu, L. C. (2017). Hayek, local information, and commanding heights: Decentralizing state-owned enterprises in China. *American Economic Review*,107(8), 2455-2478.

Levinsohn, J., & Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studie*s, 70(2), 317-341.

Olley, G. S., & Pakes, A. (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*,64(6), 1263-1297.

Syverson, C. (2004). Product substitutability and productivity dispersion. *Review of Economics and Statistics*, 86(2), 534-550.

Wang, D., Gu, J., and Yan, L. (2018). Shanghai dushiqu bianjie huafen: jiyu shouji xinling shuju de tansuo [Delimiting the Shanghai metropolitan area: using mobile phone data]. *Dili xuebao*, 10, 1896-1909.

Yu, M. (2015). Processing trade, tariff reductions and firm productivity: Evidence from Chinese firms. *The Economic Journal*, 125(585), 943-988.

Zhao, P., Hu, H., Hai X., Huang, S. & Lyu, D. (2019). Jiyu shouji xinling shuju de chengshiqun diqu dushiquan kongjian fanwei duoweishibie: yi jingjinji weili [Identifying metropolitan edge in city clusters region using mobile phone data: a case study of Jing-Jin-Ji]. *Chengshi fazhan yanjiu*, 9, 69-79+2.

Table B1: Comparison of Sample Coverage with China Statistical Yearbook and Brandt et al. (2014)

Year	Source	Number of firms	Value added	Sales	Output	Employment	Export	Net value of fixed assets
1998	Our data	165,118	1.94	6.41	6.77	61.96	1.08	4.41
	NBS	165,080	1.94	6.41	6.77	61.96	1.08	4.41
	Brandt et al. (2014)	165,118	1.94	6.41	6.77	56.44	1.08	4.41
1999	Our data	162,033	2.16	6.99	7.27	58.05	1.15	4.73
	NBS	162,033	2.16	6.99	7.27	58.05	1.15	4.73
	Brandt et al. (2014)	162,033	2.16	6.99	7.27	58.05	1.16	4.73
2000	Our data	162,887	2.54	8.42	8.57	55.59	1.46	5.18
	NBS	162,885	2.54	8.42	8.57	55.59	1.46	5.18
	Brandt et al. (2014)	162,883	2.54	8.42	8.57	53.68	1.46	5.18
2001	Our data	171,256	2.83	9.37	9.54	54.41	1.62	5.54
	NBS	171,256	2.83	9.37	9.54	54.41	1.62	5.54
	Brandt et al. (2014)	169,030	2.79	9.24	9.41	52.97	1.61	5.45
2002	Our data	181,557	3.30	10.95	11.08	55.21	2.01	5.95
	NBS	181,557	3.30	10.95	11.08	55.21	2.01	5.95
	Brandt et al. (2014)	181,557	3.30	10.95	11.08	55.21	2.01	5.95
2003	Our data	196,222	4.20	14.32	14.23	57.49	2.69	6.61
	NBS	196,222	4.20	14.32	14.23	57.49	2.69	6.61
	Brandt et al. (2014)	196,222	4.20	14.32	14.23	57.49	2.69	6.61
2004	Our data	276,474	5.72	19.78	20.17	66.22	4.05	7.97
	NBS	276,474	5.48	18.78	20.17	66.22	4.05	7.38
	Brandt et al. (2014)	279,092	6.62	20.43	20.16	66.27	4.05	7.97
2005	Our data	271,835	7.22	24.69	25.16	68.96	4.77	8.95
	NBS	271,835	7.21	24.46	25.16	67.85	4.77	8.81
	Brandt et al. (2014)	271,835	7.22	24.69	25.16	68.96	4.77	8.95
2006	Our data	301,961	9.11	31.42	31.66	73.58	6.05	10.58
	NBS	301,961	9.11	31.36	31.66	73.58	5.96	10.58
	Brandt et al. (2014)	301,961	9.11	31.36	31.66	73.58	6.05	10.58
2007	Our data	336,768	11.70	40.06	40.51	78.75	7.34	12.34
	NBS	336,768	11.70	39.97	40.52	78.75	7.31	12.34
	Brandt et al. (2014)	336,768	11.70	39.97	40.52	78.75	7.34	12.34

Note: our data is computed by summing all firms in the firm-level data; NBS's data are from China Statistical Yearbook, China Statistical Abstract, and China Industry Economy Statistical Yearbook; the unit of employment is million, and the unit of value added, sales, output, export, and net value of fixed assets are trillion yuan.

Table B2: Comparison of the Number of SOEs with China Statistical Yearbook

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
NBS	64,734	61,301	53,489	46,767	41,125	34,280	35,597	27,477	24,961	20,680
Our data	68,149	61,301	53,489	46,767	41,125	34,280	35,597	27,477	24,960	20,680

Note: The NBS data are from Table 13-8 of China Statistical Yearbook 2009.

Table C1: Estimation Results of the Production Function

SIC	Industry	Emp.	Capital	Obs.	SIC	Industry	Emp.	Capital	Obs.
13	Food Processing	0.5158*** (0.0224)	0.2930*** (0.0241)	117,848	28	Chemical Fibers	0.3993*** (0.0103)	0.3285*** (0.0116)	9,752
14	Food Production	0.5800*** (0.0178)	0.3596*** (0.0151)	47,090	29	Rubber Products	0.4238*** (0.0087)	0.3306*** (0.0081)	23,111
15	Beverage Production	0.5189*** (0.0165)	0.4002*** (0.0131)	32,808	30	Plastic Products	0.4455*** (0.0075)	0.3253*** (0.0081)	90,197
17	Textile	0.4417*** (0.0055)	0.2663*** (0.0051)	167,092	31	Nonmetal Mineral Products	0.3098*** (0.0048)	0.3626*** (0.0087)	166,936
18	Garments & Other Fiber Products	0.5524*** (0.0052)	0.2318*** (0.0328)	93,653	32	Smelting & Pressing of Ferrous Metals	0.4466*** (0.0103)	0.3475*** (0.0094)	46,034
19	Leather, Furs, Down & Related Products	0.5139*** (0.0064)	0.2445*** (0.0065)	46,258	33	Smelting & Pressing of Nonferrous Metals	0.4842*** (0.0100)	0.2750*** (0.0092)	33,171
20	Timber Processing, Bamboo, Cane, Palm Fiber & Straw Products	0.4905*** (0.0136)	0.2436*** (0.0157)	42,103	34	Metal Products	0.4379 (0.0085)	0.3172 (0.0085)	105,177
21	Furniture Manufacturing	0.6377*** (0.0473)	0.2151** (0.0932)	22,559	35	Machinery & Equipment Manufacturing	0.3730*** (0.0129)	0.3168*** (0.0145)	146,900
22	Papermaking & Paper Products	0.4235*** (0.0070)	0.3382*** (0.0081)	57,856	36	Special Equipment Manufacturing	0.3847*** (0.0145)	0.2850*** (0.0131)	80,742
23	Printing & Record Pressing	0.4276*** (0.0090)	0.5467*** (0.0083)	40,972	37	Transportation Equipment Manufacturing	0.5153*** (0.0149)	0.3480*** (0.0146)	92,676
24	Stationery, Educational & Sports Goods	0.5070*** (0.0236)	0.2386*** (0.0280)	25,574	39	Electric Equipment & Machinery	0.4771*** (0.0071)	0.3387*** (0.0077)	114,897
25	Petroleum Processing, Coking Products & Gas Production	0.2618*** (0.0117)	0.4831*** (0.0129)	14,869	40	Electronic & Telecommunications	0.5348*** (0.0098)	0.3298*** (0.0119)	63,146
26	Raw Chemical Materials & Chemical Products	0.3379*** (0.0083)	0.3670*** (0.0079)	140,963	41	Instruments, Meters, Cultural & Official Machinery	0.4193*** (0.0079)	0.2753*** (0.0074)	27,147
27	Medical & Pharmaceutical Products	0.4319*** (0.0062)	0.4358*** (0.0064)	40,347	42	Handicrafts and Miscellaneous Manufacturing	0.4699*** (0.0077)	0.2243*** (0.0059)	38,044

Note: Bootstrap standard errors from 100 repetitions are in the parentheses; *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.

Table E1: Cities Integrated with Each Other in 2000

City 1	City 2	City 1	City 2
Taiyuan Shi	Jinzhong Shi	Hefei Shi	Feidong Xian
Shijiazhuang Shi	Zhengding Xian	Hefei Shi	Feixi Xian
Shijiazhuang Shi	Luancheng Xian	Huaibei Shi	Suixi Xian
Shijiazhuang Shi	Gaocheng Shi	Tongling Shi	Tongling Xian
Shijiazhuang Shi	Luquan Shi	Fuzhou Shi	Changle Shi
Tangshan Shi	Fengrun Xian	Putian Shi	Putian Xian
Tangshan Shi	Fengnan Shi	Quanzhou Shi	Shishi Shi
Handan Shi	Handan Xian	Quanzhou Shi	Jinjiang Shi
Baoding Shi	Qingyuan Xian	Zhangzhou Shi	Longhai Shi
changzhi Shi	Lucheng Shi	Qingdao Shi	Jimo Shi
Linfen Shi	Xiangfen Xian	Dongying Shi	Kenli Xian
Liaoyang Shi	Liaoyang Xian	Anyang Shi	Anyang Xian
Nanjing Shi	Jiangning Xian	Xinxiang Shi	Xin Xiang Xian
Wuxi Shi	Xishan Shi	Puyang Shi	Puyang Xian
Xuzhou Shi	Tongshan Xian	Xuchang Shi	Xuchang Xian
Changzhou Shi	Wujin Shi	Luohe Shi	Yancheng Xian
Jiangyin Shi	Zhangjiagang Shi	Zhoukou Shi	Shangshui Xian
Suzhou Shi	Wujiang Shi	Guangzhou Shi	Zengcheng Shi
Suzhou Shi	Wuxian Shi	Shaoguan Shi	Qujiang Xian
Nantong Shi	Tongzhou Shi	Shantou Shi	Chenghai Shi
Huaiyin Shi	Huaiyin Xian	Foshan Shi	Gaoming Shi
Yancheng Shi	Yandu Xian	Jiangmen Shi	Xinhui Shi
Yangzhou Shi	Hanjiang Xian	Foshan Shi	Heshan Shi
Zhenjiang Shi	Dantu Xian	Maoming Shi	Dianbai Xian
Suqian Shi	Suyu Xian	Zhaoqing Shi	Gaoyao Shi
Hangzhou Shi	Xiaoshan Shi	Dongwan Shi	Boluo Xian
Hangzhou Shi	Yuhang Shi	Huizhou Shi	Huiyang Shi
Ningbo Shi	Yin Xian	Yangjiang Shi	Yangdong Xian
Cixi Shi	Yuyao Shi	Jieyang Shi	Jiedong Xian
Wenzhou Shi	Yongjia Xian	Nanning Shi	Yongning Xian
Cangnanxian	Pingyang Xian	Liuzhou Shi	Liujiang Xian
Wenzhou Shi	Ruian Shi	Guilin Shi	Lingchuan Xian
Wenzhou Shi	Leqing Shi	Wuzhou Shi	Cangwu Xian
Shaoxing Shi	Shaoxing Xian	Chengdu Shi	Shuangliu Xian
Shengzhou Shi	Xinchang Xian	Chengdu Shi	Pi Xian
Jinhua Shi	Jinhua Xian	Qujing Shi	Zhanyi Xian
Yiwu Shi	Dongyang Shi	Xi’an Shi	Changan Xian
Taizhou Shi	Wenling Shi		

Table E2: Descriptive Statistics of City Size in 2000

Region	Obs.	Mean	Std. Dev.	Min.	25th pct.	50th pct.	75th pct.	Max.
China	2,208	207,526	571,399	10,089	41,750	85,244	194,746	13,459,634
Five Provinces	426	350,432	623,540	16,767	104,955	202,222	364,412	7,547,467
Zhejiang	74	302,116	379,203	24,417	84,672	189,362	362,418	2,451,319
Jiangsu	77	400,810	448,625	68,442	215,196	269,089	412,891	3,510,887
Guangdong	98	484,004	1,060,231	36,813	110,703	231,406	381,745	7,547,467
Fujian	68	210,394	306,173	28,153	72,444	123,240	222,388	2,032,723
Shandong	109	314,917	436,128	16,767	104,955	170,270	330,269	2,720,972

Table G1: Estimated Selection Effect by Subsector in the Five Provinces during 1998-1999

SIC	Subsector	$\hat{S}$ (1)	$\hat{A}$ (2)	$\hat{D}$ (3)	R^2 (4)	Obs. in small cities	Obs. in big cities
17, 18, 19	Textile, Garments, and Leather, Down & Related Products	0.0184** (0.0079)	0.1401** (0.0635)	1.0506 (0.0333)	0.97	1,035	8,196
20, 21	Furniture and Products of Wood, Bamboo, Cane, Palm Fiber & Straw	0.0415** (0.0197)	-0.1492 (0.1777)	1.0898 (0.0758)	0.95	562	749
26	Raw Chemical Materials & Chemical Products	0.0488** (0.0198)	0.2577*** (0.0640)	0.9853 (0.0656)	0.97	452	2,385
31	Nonmetal Mineral Products	-0.0017 (0.0091)	0.3930*** (0.0382)	0.8876** (0.0440)	0.99	852	2,534
35, 36, 40	Equipment Manufacturing	0.0356* (0.0190)	-0.0202 (0.0760)	1.0864 (0.0747)	0.92	719	6,317

Note: (1) Small cities include the cities that are smaller than 0.5 million in 2000, did not have any highway within 50 km of their administrative boundaries in the end of 1999, and did not share any boundary with any city above 1 million. Big cities are larger than all the small cities we use. (2) In parentheses are standard errors computed from 100 bootstrapped replications. *, ** and *** denote the statistical significance at the level of 10%, 5% and 1%, respectively.