Unequal Returns to China's Intercity Road Network*

Simon Alder† Zheng (Michael) Song‡ Zhitao Zhu§

April 29, 2021

—preliminary—

link to most recent version

Abstract

We estimate the returns to infrastructure investments for each city-to-city link in China's road network. Using real-time GPS data from over half a million trucks, we first identify congested and uncongested links based on whether speed significantly decreases with traffic density. We then estimate the elasticity of traffic flows to the capacity of a link conditional on its congestion status. We incorporate congestion heterogeneity into a trade model with optimal route choices developed by Allen and Arkolakis (2019). Our structural estimation shows that the model can replicate the main features of traffic flow, speed and congestion in the data. The benefit of expanding the capacity of a link is inferred from the estimated model. The cost of the expansion is estimated from construction costs based on physical topography and market value of acquired land. We find that about 80% of China's intercity links are uncongested and associated with negative returns. The returns are much higher for congested links and the dispersion is generally large. While we focus on marginal local improvements in individual link capacity and do not quantify aggregate misallocation of the entire network, the large dispersion in returns across links suggests there could be misallocation of road infrastructure investment in China. To facilitate comparison, we also analyze real-time traffic flow data for highways in England. In sharp contrast to China, almost all intercity links in England are found to be congested.

*We thank seminar participants at the Chinese University of Hong Kong and Peking University for helpful comments. We also thank Lucas Risinger for excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect those of the SNB.

†Swiss National Bank and University of North Carolina at Chapel Hill

‡Chinese University of Hong Kong

§Chinese University of Hong Kong

1 Introduction

Transport infrastructure investment accounts for about 1% of GDP and is thus a substantial part of public investments.[1] Conditional on the size of road infrastructure investment, its welfare effect depends crucially on how the investment is allocated. Casual observations suggest ubiquitous misallocation of road infrastructure investment, from expanding the capacity of an uncongested segment to building a road in the middle of nowhere, and the marginal returns to investment in those projects must be very low, if not negative. But quantifying the returns to road investments is challenging. One reason is that an investment in one link can lead to a reallocation of economic activity across locations, and such spillover effects need to be taken into account when quantifying the returns to roads. Second, an investment in one link can lead to a rerouting of traffic or an increase in the demand for trips, which can generate enough additional traffic that congestion ends up being the same as before the investment (see, e.g., Duranton and Turner, 2011, on this *fundamental law of road congestion*). However, this elasticity of traffic with respect to road investments depends on the congestion status of a road, which may be heterogeneous.

The goal of this paper is to tackle these two challenges. We first use unique real-time truck GPS records to estimate congestion for each city-to-city link in China's road network. The links with or without a significantly negative correlation between high-frequency intra-day traffic flows and speeds are identified as congested or uncongested links. Strikingly, we find more than 80% of China's intercity road network to be uncongested. We then estimate the traffic elasticity from variation of capacity (i.e., lane kilometers) across links. Inspired by Duranton and Turner (2012) and Baum-Snow et al. (2020), we use China's road network in 1962 as an IV to address the potential endogeneity of intercity link capacity. The estimated elasticity is about 1.4 and 0.6 for congested and uncongested links, respectively. We also run panel regressions to estimate the responses of speed to an expansion of link capacity over time. Speed on a congested link increases significantly in the first quarter after new lanes were added in the previous year. The effect gradually diminishes and becomes insignificant in the second half of the year. In contrast, the monthly effects are always insignificant for uncongested links. The results are robust to an IV constructed to control for endogeneity in the time-series variation of link capacity. The estimates from both cross-sectional and panel regressions are not only consistent with the fundamental law of road congestion but also provide external validity for the congestion status identified by our high-frequency traffic data.

We then introduce the congestion heterogeneity into an economic geography model with

[1] See Fay et al. (2019) for a recent assessment of infrastructure expenditures in a large sample of countries. Gurara et al. (2017) document that public investment accounts for about 4-6% of GDP in low-income and emerging countries.

optimal route selection developed by Allen and Arkolakis (2019). In the model, investing in road infrastructure on a city-to-city link reduces the trade cost between the two cities through improvements in road capacity. In addition to the direct local effect, there are indirect effects on all city-pair trade costs through driving time, which may decrease or increase in response to road infrastructure investment in a specific locality, depending on the optimal responses of route choices and congestion. The impact on trade costs affects trade flows, which, in turn, affect traffic flows, route choices, and feed back into trade costs through congestion. The general equilibrium pins down congestion and traffic elasticity for each link, which determine the effect of road infrastructure investment on trade costs and, ultimately, on welfare.

The model is structurally estimated by matching the observed average traffic flows and speed on each intercity link. The simulated data by the estimated model fit the observations well. The correlation is above 0.8 for traffic flows and close to 1 for speed. More importantly, the estimated model implies the traffic elasticity of 1.3 and 0.6 for congested and uncongested links, respectively. The indirect inference shows that the estimated model can generate traffic elasticities very close to the estimates in the reduced-form regressions, which are not targeted in the structural estimation. The estimated model quantifies the elasticity of aggregate welfare to the capacity of a link. The average welfare elasticity is 0.036‰ and 0 for congested and uncongested links, respectively. The welfare elasticities turn out to be robust to alternative assumptions on the agglomeration effects and labor mobility.

To complete the estimation of marginal returns to road infrastructure investment, we also measure the cost side, which consists of construction costs, opportunity costs of acquired land, capital, and maintenance costs. The land opportunity costs are particularly important since more congested areas tend to have higher land values. The implied marginal returns to road infrastructure investment have an average of 5.2% for congested links. The average returns are negative for uncongested links. The returns also vary substantially within congested links, with a standard deviation of 0.104. For example, investing in the link between Wuxi and Suzhou has the highest return of 78.93%, while investing in the link between Chengdu and Deyang, with a similar congestion status, can only deliver a return of 20.49%.

The dispersion of marginal returns could be a manifestation of misallocation in the spirit of Restuccia and Rogerson (2008) and Hsieh and Klenow (2009), although the concept cannot be applied directly because of complementarities between different road investments. While we do not attempt to quantify the aggregate misallocation, the available evidence suggests that the large dispersion in returns when investing in each individual link is indeed likely a sign of misallocation. One reason is that the observed expansion of road capacity was uncorrelated with the estimated returns across links. In fact, uncongested links expanded at a slightly faster pace than congested links. Another reason is that the highly unequal returns are also hard to reconcile with the observed spatial productivity growth. We first calculate

returns to road investment in a model calibrated to the Chinese economy in 2014. We then conduct a counterfactual in which city productivity is increased to its level in 2018. The productivity growth increases returns to road investment by one third in the counterfactual exercise. However, the dispersion of the returns is barely affected.

A takeaway from the paper is that about 80% of China's intercity links are uncongested and associated with nearly zero or even negative marginal returns. However, there may be factors omitted from the model that cause different marginal returns across links.[2] As a comparison, we analyze real-time traffic flow data for highways in England using the same empirical methods. In contrast to China, almost all intercity links in England are found to be congested. To make a more appropriate comparison, we construct a restricted version of China's road network that only covers national- and provincial-level roads (which are more comparable to highways in England) and three coastal regions (with an income per capita more comparable to that in England). The contrast remains sharp. Uncongested links account for about two thirds of the total capacity in the restricted road network in China.

The remainder of the paper is organized as follows. In the rest of this section, we review the related literature. Section 2 describes the data on cities, roads, and traffic. Section 3 discusses congestion and how we measure it. Section 4 introduces the economic geography model. Section 5 discusses the structural estimation. Section 6 describes the welfare gains from infrastructure investments along different links of the network. Section 7 concludes.

1.1 Literature

The paper contributes to a growing literature that quantifies the general equilibrium effects of inter-city transportation infrastructure (e.g., Allen and Arkolakis, 2014; Donaldson and Hornbeck, 2016; Donaldson, 2018). We contribute to this literature by using truck-level GPS data to measure the speed and density of traffic throughout the network and identifying congestion. To incorporate congestion and optimal route selection into the general equilibrium model, we follow Allen and Arkolakis (2019). We differ from their approach by allowing the parameters that determine congestion to differ across links and showing that this congestion heterogeneity is important for the welfare effects. The analysis of congestion is also closely related to the literature on the relationship between road investments and congestion (e.g., Duranton and Turner, 2011; Hsu and Zhang, 2014; Chen and Klaiber, 2020; Garcia-López et al., 2020).

Our finding that there is a large dispersion in the returns to road investments across links in the national highway network is in line with the results in Allen and Arkolakis

[2] Measurement errors, adjustment costs and markup variations are among the factors that tend to overestimate misallocation of production inputs and have been intensively studied in the recent literature (Asker et al., 2014; Song and Wu, 2015; Bils et al., 2020).

(2019). Our findings also relate to studies that use general equilibrium models to quantify the misallocation of transportation infrastructure investments (Alder, 2019; Alder and Kondo, 2019; Fajgelbaum and Schaal, 2020) and the misallocation literature more broadly (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009), but we do not quantify the aggregate cost of misallocation.

Several existing studies have proposed strategies to identify the causal effects of transportation infrastructure when there are concerns that the infrastructure investments are endogenous (e.g., Chandra and Thompson, 2000; Baum-Snow, 2007; Duranton and Turner, 2012). We rely on these strategies in the cross-sectional analysis and propose a novel time-varying instrument for the panel analysis.

A number of papers have studied transportation infrastructure investments in China (e.g., Faber, 2014; Banerjee et al., 2020; Baum-Snow et al., 2020; He et al., 2020) and we contribute to this literature by estimating congestion using new truck-level GPS data and documenting the heterogeneity in returns across links. Truck-level data is also used in Allen et al. (2020b), but their focus is on market power, and they do not use GPS data on trucks' location. Krishna and van Leemput (2018) use truck-level data in India to study internal trade barriers between regions, while we focus on the returns to road capacity investments on links along the national highway network in China.

2 Data

We combine detailed data on Chinese cities and roads to construct a national road transportation network and then connect it to the spatio-temporal big data based on GPS devices in 562,980 trucks in China. Furthermore, we use several socio-economic datasets such as GDP, population, and night lights.

2.1 Cities and Roads

We treat each prefecture as a "city" or a node in the network and obtain the geographic information using the map from China's official source, "National Catalogue Service for Geographic Information".[3] This map provides each city's administrative boundary and we compute the corresponding geometric center using ArcGIS. We exclude islands (Zhoushan and Hainan) and north-western areas (Xinjiang, Gansu, Ningxia, Qinghai, Tibet and Inner Mongolia), where the truck business is underrepresented for the company who provides us with the GPS data, and thus we have 273 cities in total.

[3]Beijing, Tianjin, Shanghai, and Chongqing, which are provincial level cities, are also included.

The roads connecting cities also come from the "National Catalogue Service for Geographic Information." It contains national, provincial, county, and town level roads, as well as those within cities and villages, and we mainly focus on national, provincial, county, and within-city roads (abbreviated as "NPCC"). In other words, we drop town level roads, roads for specific purposes, and roads within counties, because these smaller roads typically do not have enough truck traffic. Throughout the paper, we use the equidistant projection *WGS 1984 World Equidistant Cylindrical*.

2.2 Road Network

Following the conventional approach in the existing literature, we construct a network connecting cities through the shortest path along roads, which is computed with "closest facility analysis" in the "Network Analyst" toolbox in ArcGIS. Specifically, in the first step of constructing the road network, we select the lowest cost routes (in terms of distance) from each city to all its neighbors. In this case, for two neighboring cities, like Shanghai and Suzhou, drivers can move directly from one city to another or indirectly (e.g., from Shanghai to Jiaxing and then to Suzhou) according to their idiosyncratic preferences as modeled in the quantitative analysis. Between two cities that are not neighbors, such as Shanghai and Nanjing, drivers can only move indirectly from one city to another (e.g., through Suzhou, Wuxi, or Changzhou), but different drivers might choose different paths.

Furthermore, instead of using only *one shortest path* between the prefecture-level cities' centroids, we connect two neighboring cities using a *group of shortest paths* among the counties within the prefectures. This group of shortest paths, which we refer to as a link, consists of all shortest paths from each county in one city to each county in the other city. For example, the link from city A (including county centroids a_1, a_2) to B (including county centroids b_1, b_2, b_3) consists of six connections: from a_1 to b_1, from a_1 to b_2, from a_1 to b_3, from a_2 to b_1, from a_2 to b_2, and from a_2 to b_3. The shortest paths connecting two county centroids are constructed in the same way as we construct the shortest paths connecting two cities. In this case, more road segments can be covered in the network, and thus more truck information is included in our analysis. The entire network is illustrated as orange lines in the left map in Figure 1 and the map on the right shows as an example the paths from Shanghai to Suzhou and Jiaxing.

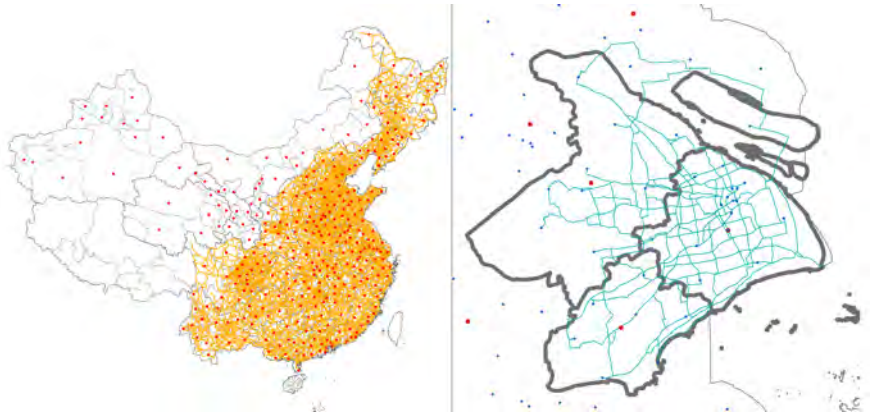


Figure 1: Road Network

Note: In the map on the left, the orange lines show the current network and the red points denote cities. In the map on the right, we illustrate the road links between Shanghai and its two neighbors, Suzhou and Jiaxing.

For each shortest path, we obtain the lane information from the navigation road map from the Geographic Data Sharing Infrastructure.[4] We spatially match the navigation map with the official road map to obtain the lane information of each road segment. Then we compute the average lane number across all segments along the shortest path by using each segment's length as the weight. Analogously, the weighted average is also used to calculate the lane number and speed limit for each link.

Based on this network, we introduce some key definitions and variables.

Definition 1 (County-pair shortest path)**.** The shortest path connecting a county in a city to another county in a different but neighboring city. County-pair shortest paths have directions.

Definition 2 (Capacity of a county-pair shortest path)**.** The length of the shortest path multiplied by the average number of lanes on the shortest path.

Definition 3 (City-pair link)**.** The set of all county-pair shortest paths in two neighboring cities. City-pair links also have directions. We denote by (k, l) the city-pair link from city k to l.

A city-pair link consists of all county-pair shortest paths and, therefore, differs from the city-pair shortest paths in Allen and Arkolakis (2019), which focus on the shortest paths between cities. However, the difference might not be large because Allen and Arkolakis (2019) use traffic flows on the Interstate Highway System between cities, which are arguably much more likely to go through city-pair shortest paths. Our focus is on city-to-city truck flows. In fact, 30.71% of the trucks were on county-pair shortest paths, while only 6.36% were on

[4]The Geographic Data Sharing Infrastructure is maintained by the College of Urban and Environmental Science, Peking University, see http://geodata.pku.edu.cn.

city-pair shortest paths. Our city-pair link based on county-pair shortest paths provides a much broader coverage of city-to-city truck flows.

The capacity and length of a city-pair link are defined as follows.

Definition 4 (The capacity of a city-pair link: cap_{kl})**.** The sum of the capacity of all county-pair shortest paths in two neighboring cities.

Definition 5 (The length of a city-pair link: $dist_{kl}$)**.** The average length of county-pair shortest paths in two neighboring cities.

The capacity of a city-pair link may increase mechanically in the number of county pairs that the city pair has, and the correlation is as high as 0.85. We will control for the number of county pairs in a city pair in the following cross-sectional regressions. The length of a city-level link corresponds to the driving distance between two neighboring cities. It is slightly longer than but highly correlated with the length of the straight line connecting the two city centers (see Figure 11).

2.3 Trucks

Trucking is the primary mode of domestic freight transport in China. Highway, for example, accounts for 73% of the total freight in China in 2019 by official statistics. We obtained real-time truck GPS data from G7, one of China's leading logistical service providers. The data covers 562,980 heavy trucks between January 2013 and December 2018, accounting for around 7.86% of all heavy trucks in China in 2018. Since heavy trucks are primarily for shipment of goods across cities, we argue that truck flows indicated by our GPS data mainly capture city-to-city goods flows.

There are obvious concerns on the representativeness of the data. However, using the same GPS data aggregated to city pairs, Chen et al. (2020) find truck flows associated with a city to be strongly correlated with city-level GDP (correlation 0.90) and night light intensity (correlation 0.86). The correlations in our sample are even stronger.[5] Another issue is the growing share of heavy trucks covered by the data (from 0.37% in 2013 to 7.86% in 2018) due to the rapid expansion of the company, which might bias our estimates. In what follows, we will mainly rely on the 2018 GPS records. In some empirical exercises where earlier data are necessary, we will use truck speed only, which is less likely to be biased by the changing market shares of the company.

We obtain the following real-time GPS records for all trucks in the data. For each month, our data cover two days (one weekday and one weekend in the middle of each month); and

[5] Chen et al. (2020) includes all the prefectures excluding those in Tibet and Xinjiang with much fewer trade linkages with the rest of the country.

for each day, our data cover four hours (8 a.m. – 9 a.m., 12 a.m. – 1 p.m., 5 p.m. – 6 p.m., 9 p.m. – 10 p.m.). We have GPS signals about every 10 seconds and with a precision of 1e-5 degrees. The data include the latitude, longitude, and speed for each truck, as well as the time and date of the observation.

The truck data is cleaned through the following four steps: (1) we drop records where speed is greater than 120 km/h (0.005%), which is the speed limit of China's highway, or negative; (2) we drop records where a truck keeps zero speed for at least 15 minutes (39.370%), because such cases could be due to parking instead of congestion; (3) we drop records where a truck's latitude or longitude moves greater than 1 degree within several seconds (0.027%); (4) we drop trucks whose traces' lengths are greater than 120 km (0.001%), because we focus on intervals of one hour and the speed limit is 120km/h.

We then assign each observation to a location on a digitized map of China according to the longitude and latitude. We match these points to specific roads according to the distance. Specifically, for each road, we code all trucks that are at most 30 meters away from it as driving on the road. In addition, by tracing each truck's exact location over time, we can identify the truck's driving direction, which enables us to select trucks and roads that share the same directions.

Based on the data preparation procedure above, we obtain a comprehensive picture of China's trucks at any snapshot in our sample period. In practice, we treat each one minute as a snapshot,[6] and construct several variables for the empirical and quantitative exercises later on. The following variables can be directly constructed from our data:

$speed_{kl,t}$: the average speed over all trucks on link kl within one minute t;

$traf_{kl,t}$: the number of different truck IDs on link kl within one minute t.

Some related variables can also be derived from these two:

$hours_{kl,t} \equiv dist_{kl}/speed_{kl,t}$: the driving hours from city k to l along link kl, and it is specific to the same minute t for which we measure speed;

$density_{kl,t} \equiv traf_{kl,t}/cap_{kl}$: the number of trucks per unit of area (1 mile $\times$ 1 lane) on link kl, and it is also specific to the same minute t for which we measure $traf_{kl}$.

In addition to the congestion estimation in Section 3.1, where we exploit across-minute variation within a year, the other analyses focus on either the cross-sectional variation (Sections 3.2.1, 5, and 6) or across-year variation (Section A.3.3). Therefore, we aggregate over minutes within each year and obtain the following variables:

[6] Although most trucks report their GPS position every ten second, there are still some trucks with longer gaps. Thus, we select one minute to avoid missing trucks with such longer reporting gaps as much as possible.

$speed_{kl}$: the mean value $speed_{kl,t}$ across all minutes within a year ;

$traf_{kl}$: the mean value of $traf_{kl,t}$ across all minutes within a year.

The remaining two variables can be constructed correspondingly as

$hours_{kl} \equiv dist_{kl}/speed_{kl}$: the average driving hours from city k to l along link kl;

$density_{kl} \equiv traf_{kl}/cap_{kl}$: the average number of trucks per unit of area (1 mile $\times$ 1 lane) on link kl.

All the above definitions and variables treat each link as one. In the next step we will zoom into the detailed components of each link, the points, to prepare for the introduction of the concept of traffic flow or volume, which will be defined in Equation (1) below. Here, we make the assumption that trucks are homogeneously distributed within each link,[7] and then the above $speed_{kl}$, which is the average speed across all points on link kl, can also be treated as the speed at any point on link kl. Hence, the number of trucks per unit of distance ($density_{kl} \times lane_{kl}$) can also be applied to any sub-set of link kl.[8]

Following Small and Verhoef (2007), we calculate the traffic flow or volume $flow_{kl}$, defined as the number of trucks passing a given point on link kl per unit of time, as the product of density, speed and lane number

$$flow_{kl} \equiv speed_{kl} \times density_{kl} \times lane_{kl}, \tag{1}$$

which is consistent with the units of measure: $\frac{number}{hours} = \frac{number}{mile \times lane\ number} \times \frac{mile}{hours} \times lane\ number$. Under the assumption of within-link homogeneity, $flow_{kl}$ can be treated as the number of trucks passing *any point* on link kl per unit of time. In addition, $flow_{kl}$ measured in different units of time are proportional to each other (e.g., $flow_{kl}$ per hour is proportional to $flow_{kl}$ per year), and thus we can infer number of trucks passing any point on link kl per year from $flow_{kl}$.

This variable as well as the two assumptions will be useful in two ways: First, the definition of $flow_{kl}$ implies that this variable is analogous to traffic flows captured by traffic cameras, and the later is widely used in the existing literature examining the fundamental law of road congestion.[9] Thus, this $flow_{kl}$ can be applied to construct the dependent variable, vehicle-mile

[7] In other words, here we abstract from the heterogeneity within each link, but we will discuss this aspect at the end of Section 3.2.1.

[8] This assumption does not require the number of lanes to be homogeneous across all segments of a link. In fact, the variance of the number of lanes within a link is exploited to construct a new instrument in Section A.3.3.

[9] For example, it is similar to "number of vehicles per lane per day passing any point", called "average annual daily traffic (AADT)" in Duranton and Turner (2011), "the average number of vehicles passing through the

travelled, in our analysis of the fundamental law of road congestion. Second, this variable can serve as a counterpart of the total trade flow through link kl, Ξ_{kl}, which is defined as the total value of goods moving through link kl in a year. Given the two assumptions, we can recover from $flow_{kl}$ the number of trucks passing *any point* on link kl *per year*, which can approximate the total number of trucks moving through link kl in a year, and thus becomes a natural counterpart to Ξ_{kl}.[10]

2.4 Socioeconomic Data

For each city, we obtain GDP and population data from *China's Regional Statistical Yearbook* and *China's Urban Statistical Yearbook*. We also use night light data as well as grid-level GDP and population data, which allow us to evaluate the socioeconomic conditions along the link. The night light data are provided by Flint from China Remote Sensing Satellite Ground Station, Chinese Academy of Science, which is constructed based on the original data from *Version 1 VIIRS Day/Night Band Nighttime Lights* for 2018 with a grid size of 15 arc-second.[11]

3 Empirical Analysis of Congestion

Our GPS data includes information on the location and speed of trucks, which we can use to measure the overall density of trucks on a given road segment and at a given time. The combination of speed and traffic density allows us to estimate a standard congestion equation and identify congested links in the network. Using variation in link capacity across roads and years, we also test the fundamental law of road congestion and compare congested and uncongested roads.

3.1 Measuring Congestion

In order to characterize the spatial distribution of congestion, we use a congestion equation originally proposed by Vickrey (1967) to capture the relationship between vehicle density

segment observation point between 7 am and 7 pm on a weekday", called "average weekday daytime traffic (AWDT)'' in Hsu and Zhang (2014), and the number of cars passing any noise monitor on the urban road system in 20 minutes in Chen and Klaiber (2020).

[10]To avoid notation redundancy, we omit the scale parameter that transforms traffic flow per hour to that per year, and directly use $flow_{kl}$. The only minor caveat is that the scale parameter affects the interpretation of ζ, but it would not affect the validity of welfare analysis.

[11]The night light data are available at http://satsee.radi.ac.cn/cfimage/nightlight/index_en.html, which are based on https://ngdc.noaa.gov/eog/download.html.

and driving speed and then empirically estimate this equation for each link.[12] We modify the specification in Vickrey (1967) and Allen and Arkolakis (2019) to accommodate the capacity of city-pair links:

$$\frac{1}{speed_{klt}} = \frac{1}{\delta_{0kl}} + \delta_{1kl}\left(\frac{traf_{klt}}{cap_{kl}}\right)^{\delta_{2kl}}, \tag{2}$$

where $traf_{klt}$ denotes the number of trucks on city-pair link (k,l) at time t, $speed_{klt}$ is the average speed of trucks on (k,l) at time t. In Vickrey (1967), the left-hand variable is $hours_{klt}/dist_{kl}$, where $hours_{klt}$ denotes the travel duration from k to l at time t. Note that $1/speed_{klt}$ is equal to $hours_{klt}/dist_{kl}$. We use speed as it is directly observed in the data. δ_{0kl} can be interpreted as the uncongested speed.

To estimate the congestion equation, we rewrite (2) as

$$\ln\left(\frac{1}{speed_{klt}} - \frac{1}{\delta_{0kl}}\right) = \ln\delta_{1kl} + \delta_{2kl} \times \ln\left(\frac{traf_{klt}}{cap_{kl}}\right) + \epsilon_{klt}. \tag{3}$$

We identify δ_{1kl} and δ_{2kl} from within-link speed variation over time by estimating the parameters for each link separately. Specifically, we compute the average truck speed and total number of trucks minute by minute in the four time slots 8:00-9:00, 12:00-13:00, 17:00-18:00 and 21:00-22:00. We select four minutes (15, 30, 45, 60) in each hour.[13] Since our data covers one working day and one weekend day in each month of 2018, there are a total of 384 observations of $speed_{klt}$ and $traf_{klt}$ for each link. $dist_{kl}$ and cap_{kl} are obtained from China's navigation data. δ_{0kl}, the uncongested speed, is proxied by maximal value of observed speeds.[14]. We use OLS to estimate δ_{1kl} and δ_{2kl} for each link separately.[15]

The estimated δ_2 is positive and statistically significant at the level of 5% for 307 out of 1,464 city-pair links.[16] We identify the links with significantly positive δ_2 as congested links and refer to the other links as uncongested. The mean estimate of δ_2 for congested links is 0.236, with a standard deviation of 0.156. We plot the spatial distribution of the estimated δ_2 in the left panel of Figure 2, where the red lines represent congested links. The distribution of significantly positive δ_2 are plotted in the right panel.

[12] In the structural estimation in Section 5, we will directly link trade costs to driving hours (trip duration), which we can infer from driving speed.

[13] The selection is to make the data structure comparable to that of England as discussed in Appendix A.2.3.

[14] If such speed exceeds the average speed limit of the link (weighted by the length of each road in the link), we would use the speed limit instead. The results are robust when we use speed limit to proxy δ_{0kl} for all links.

[15] One potential concern is that many trucks in China would be so overloaded that the variation of driving speed would negligible. However, as illustrated in Figure 9 and 10, our data tell there are huge variations for us to estimate congestion validly.

[16] The estimated δ_2 is positive and significant at the level of 10% for 418 links.

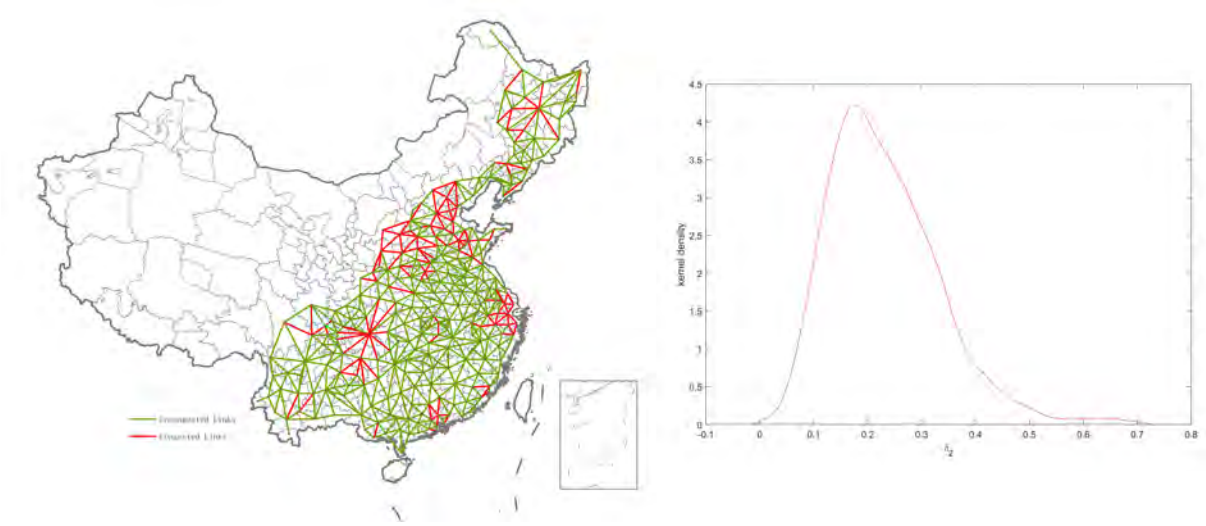


Figure 2: Distribution of δ_{2ij}

Note: In the left panel, the red lines indicate congested links with significant and positive δ_{2ij}, the green lines indicate uncongested links. In the right panel, we focus on the positive and significant δ_2s.

Note that our approach differs from Allen and Arkolakis (2019), who exploit cross-link variation and assume δ_0, δ_1 and δ_2 to be the same for all links. In contrast, our estimated δ_0, δ_1 and δ_2 are all *link-specific*. Since the level of congestion turns out to be very different across links, allowing the parameters to be link-specific and distinguish congested and uncongested links is an important extension for the quantitative exercise.

We also compare our results to congestion patterns in England. Appendix A.2.3 provides further details and we briefly discuss the main results here. England has detailed traffic data from more than 7000 measurement sites along its highway network. Importantly, the English data include information on speed and traffic counts for every 15-minute interval. This allows us to study the relationship between traffic and speed similar to our analysis for China, although the origin of the data (from fixed measurement sites) is different from the Chinese data (based on GPS devices in trucks). While the English traffic data is detailed and allows for a similar empirical analysis as in China, an important difference between England and China (and many European countries) is that England spent less on road infrastructure over the past decades, and its traffic density is relatively high.[17] While these differences are, on the one hand, a challenge for the comparison between England and China, they are, on the other hand, also an opportunity to consider potential long-term dynamics when China moves towards the steady state with higher traffic density and lower investments (see Section 6.6).

[17]For a comparison of England's network to European networks, see e.g. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/212590/action-for-roads.pdf.

Overall, the density of highways in China is substantially lower than in England, but it is more comparable when we restrict the sample to the denser Yangtze River Delta, Pearl River Delta, and Beijing-Tianjin-Hebei regions. We estimate the congestion equation (3) for highways in China and England. The spatial distribution of the estimated δ_2 (analogous to the chart on the left in Figure 2) is plotted in Figure 15 in the appendix. The density of positive and significant δ_2 (analogous to the chart on the right in Figure 2) is plotted in Figure 16 in the appendix. Overall, nearly all links in England (120 out of 122) are congested. The share of congested links in China's highway network is 41%. While the share is higher than 20.97% in our benchmark Chinese road network, it is still much lower than the congested share in England.

3.2 The Fundamental Law of Road Congestion

The estimation of the congestion equation leads to a sharp contrast between China and England on the proportion of congested links. In this section, we empirically test the fundamental law of road congestion on the congested and uncongested links identified by the estimated congestion equation. The empirical exercise provides a validity check of our congestion measure. If the identification of congestion is valid, we would expect the fundamental law of road congestion to hold for the congested links but to fail for the uncongested ones. In what follows, we will first use cross-sectional variations in city-pair link capacity and run OLS regressions. Both endogeneity of capacity across city-pair links and heterogeneity of roads in a city-pair link can bias the OLS estimate. We address the two challenges by constructing an IV based on historical roads and estimating road-level congestion.[18]

We then use panel data to explore variations in city-pair link capacity and traffic speed over time. To account for the potential endogeneity in the development of city-pair link capacity, we construct a novel IV. The high-frequency of GPS data allows us to capture the dynamic response of truck flows to an expansion of link capacity. While the prediction of the fundamental law on speed applies to both congested and uncongestged links in the long run, we find a significant short-run speed increase on congested links in response to an expansion of link capacity. Such a short-run effect doesn't exist for uncongested links.

3.2.1 Cross-Sectional Regressions

We first run OLS regressions using cross-sectional variation in link capacity based on the following specification:

$$\ln(VMT_{kl}) = \gamma_1 + \gamma_2 \times \ln(cap_{kl,-1}) + \rho \times Z'_{kl} + \epsilon_{kl}, \tag{4}$$

[18]Each link consists of several roads, where a road is defined by the official name.

The dependent variable, VMT_{kl}, is vehicle-mile-travelled in 2018 and the main independent variable, $cap_{kl,-1}$, is link capacity at the end of 2017.[19] We use the GPS data only for 2018 because the number of trucks in the data increased dramatically in 2013-17 and reached a relatively stable level in 2018. Z'_{kl} contains a series of control variables, including the average speed limit on a link, the average nightlight density and ruggedness along a link,[20] and the total number of counties within a city pair.[21] The average speed limit captures road quality of the link, the average nightlight density measures the density of economic activity, and the ruggedness proxies the geographic endowment. We also add the number of county pairs in a city pair to control for the mechanical effect of the number of shortest paths on the city-pair link capacity.

The results are presented in Table 1. Column (1) reports the results from all links. The estimated elasticity of vehicle-mile-travelled with respect to link capacity is 1.05. Column (2) and (3) separate the sample into congested and uncongested links. The estimated elasticity for congested links is a fifth larger than that for uncongested links and the difference is statistically significant at the 10% level.

[19] 2017 is the most recent year for which we can access the lane-level navigation road map. One-year lag may also mitigate reverse causality.

[20] We average nightlights across grids in the 10km buffers along a link. The 1km buffers are dropped because road emissions from the roads may intensify nightlights within the buffers.

[21] Two historical control variables (population density in 1964 and urban land share in 1982) which are used in our IV estimation are also controlled in Table 11 to illustrate the robustness of the results.

Table 1: OLS

Dep: VKT $(\ln)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Capacity $(\ln, lag)$	1.076***	1.304***	1.041***
	(0.057)	(0.101)	(0.068)
Speed limit $(\ln, lag)$	3.625***	1.513***	4.072***
	(0.203)	(0.410)	(0.238)
Night light $(\ln)$	1.083***	1.991***	0.978***
	(0.081)	(0.186)	(0.101)
Ruggedness	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)
Number of county pairs	0.000	0.000	0.000
	(0.000)	(0.000)	(0.001)
N	1464	307	1157
R^2	0.593	0.679	0.563
p value for H_0		0.038	

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. H_0: the coefficients of Capacity $(\ln, lag)$ in congested and uncongested groups are the same.

There are several sources of endogeneity that may bias the estimated elasticity. The correlations can be driven by reverse causality, i.e., link capacity endogenously responds to VMT and demand factors. To estimate the causal effect of road capacity on congestion, the transportation infrastructure literature has developed several instrumental variables. Following Duranton and Turner (2011), we construct an IV from the 1962 road network as in Baum-Snow et al. (2020). Specifically, the IV is the total road length in a city pair in 1962. The relevance of this instrument relies on the argument that constructing new roads alongside or even on top of the existing roads is less expensive. The first stage confirms that historical roads are a strong predictor of today's roads. The exogeneity of this instrument requires that the historical roads cannot be correlated with today's traffic through other channels than today's road capacity. Baum-Snow et al. (2020) discuss in detail that the historical roads in 1962 are built under an agricultural and planned economy, and with the purpose of connecting rural and cities, rather than constructing inter-city links. Furthermore, we directly control for population in 1964 to capture the historical economic conditions, and urban land share

in 1980 to proxy the historical industrial structure.[22] To further alleviate the concern from unobservable factors, which may be serially correlated and predict roads in 1962 as well as today's flows within the same city pair, we also follow the approach in Baum-Snow et al. (2020) to refine our instrument by first calculating total length of historical roads in a city pair with a buffer of 10 kilometers and then excluding the length within the city pair.[23]

The 2SLS regression is conducted based on the following specification:

$$\ln(cap_{kl,-1}) = d_1 + d_2 \times \ln(hist_{kl}) + \varrho \times Z'_{kl} + \varepsilon_{kl}, \tag{5}$$

$$\ln(VMT_{kl}) = \gamma_1 + \gamma_2 \times \ln(\widehat{cap_{kl,-1}}) + \rho \times Z'_{kl} + \epsilon_{kl}. \tag{6}$$

Table 2 reports the results for the first stage. Our IV is constructed to capture supply-side effects on the current link capacity. We find a highly significant positive correlation between the historical road length in a city pair and the current capacity of the city-pair link. This is consistent with the finding in the existing literature (e.g., Duranton and Turner, 2011; Baum-Snow et al., 2020). The F statistic is substantially above 10, suggesting a strong first stage.

[22] We also find that conditional on the population in 1964, historical road length in 1962 is not correlated with later population growth between 1964 and 2018. The population data for 1964 is obtained from *The Compilation of China's Population Statistic Materials (1949-1985)* and 1964 is the year closest to 1962 when we can access city level population data. The urban land share in 1980 is obtained from *The Dataset of China's Land Use from Remote Sensing (1980)* at http://www.resdc.cn/data.aspx?DATAID=197, which is also the earliest land use data available to us.

[23] The results are robust when we apply a wider buffering distance like 50 kilometers (see Table 12 and 13).

Table 2: 2SLS Results, First Stage

Dep: Capacity $(\ln, lag)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Historical road length in 1962 $(\ln)$	0.244***	0.236***	0.219***
	(0.029)	(0.044)	(0.033)
Speed limit $(\ln, lag)$	0.194	-0.910***	0.318
	(0.177)	(0.246)	(0.194)
Night light $(\ln)$	-0.303***	-0.201	-0.332***
	(0.056)	(0.124)	(0.063)
Population in 1964 $(\ln)$	0.250***	0.206***	0.273***
	(0.028)	(0.035)	(0.033)
Urban land share in 1980	7.525***	5.067	4.585*
	(2.346)	(3.985)	(2.553)
Ruggedness index	0.083***	0.024*	0.083***
	(0.013)	(0.014)	(0.016)
Number of county pairs	0.004***	0.003***	0.005***
	(0.000)	(0.000)	(0.000)
N	1464	307	1157
R^2	0.675	0.744	0.664
Kleibergen-Paap rk Wald F statistic	68.759	28.344	43.721

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. The IV is the total length of historical road in a buffer of 10 kilometers extended from the city pair (excluding the length within the city pair)

Table 3: 2SLS Results, Second Stage

Dep: VMT $(\ln)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Capacity $(\ln, lag)$	0.735***	1.440***	0.581***
	(0.167)	(0.369)	(0.218)
Speed limit $(\ln, lag)$	3.242***	0.605	3.799***
	(0.198)	(0.493)	(0.222)
Night light $(\ln)$	0.404***	2.006***	0.249*
	(0.122)	(0.339)	(0.145)
Population in 1964 $(\ln)$	0.537***	0.375***	0.605***
	(0.064)	(0.108)	(0.084)
Urban land share in 1980	25.347***	-6.419	29.008***
	(3.597)	(6.851)	(4.219)
Ruggedness	0.126***	0.067***	0.162***
	(0.020)	(0.026)	(0.026)
Number of county pairs	-0.000	-0.002	0.001
	(0.001)	(0.001)	(0.001)
N	1464	307	1157
R^2	0.634	0.708	0.603
p value for H_0		0.010	

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. The IV is the total length of historical road in a buffer of 10 kilometers extended from the city pair (excluding the length within the city pair). H_0: the coefficients of Capacity $(\ln, lag)$ in congested and uncongested groups are the same.

Table 3 reports results for the second stage. We find a much larger difference of the estimated elasticity between congested and uncongested links. The estimate for congested links almost triples that for uncongested links. This contrast also highlights the importance of congestion heterogeneity. More importantly, the estimated elasticity for uncongested links is substantially below one, which violates the fundamental law of road congestion that "predicts that the elasticity of traffic to road capacity is at least 1" (Hsu and Zhang, 2014).

Another potential bias is that congestion is identified at the level of city-pair links, but there exists rich heterogeneity of congestion status across roads within each link. To see this, we estimate exactly the same congestion equation at the road level and use the same criteria to identify congested roads by the road-specific estimate of δ_2. Figure 17 in the appendix

plots the distribution of the proportion of congested roads for congested links (solid line) and uncongested links (dotted line), weighted by road capacity. The average share of congested roads for congested links is indeed higher than that for uncongested links (19.01% v.s. 13.23%). This is consistent with our link-level congestion measure that captures the average congestion status. Furthermore, we find substantial heterogeneity in the proportion of congested roads across congested or uncongested links. To assess the potential bias caused by within-link road-level congestion heterogeneity, we select the longest provincial or national road in each link. The OLS and IV results are reported in Table 15 in the appendix. The estimated elasticity for uncongested roads is about 40% less than that for congested roads. The significant difference of estimates between congested and uncongested groups from road-level variation is similar to that from link-level variation. We conclude that our findings are qualitatively robust to within-link congestion heterogeneity.

We also run panel regressions to estimate the responses of speed to an expansion of link capacity in Appendix . We find significant differences in the responses between congested and uncongested links. The results are robust to using an IV based on supply-driven bottlenecks in the networks over time to control for endogeneity in the time-series variation of link capacity.

4 Model

We begin with an economic geography model as in Allen and Arkolakis (2014). We then introduce optimal route selection to endogenize road congestion *à la* Allen and Arkolakis (2019). Our goal is to build a framework that allows us to quantify the impact of road infrastructure on aggregate welfare.

4.1 An Economic Geography Model

We consider a discrete version of Allen and Arkolakis (2014), where the economy consists of a finite number of locations indexed by $i \in \{1, \cdots, N\}$. The total number of agents is constant and denoted by $\bar{L}$. The benchmark model assumes perfect labor mobility, i.e., the agents can freely move across locations.

The preferences of a representative agent in location j are

$$U_j = \left(\sum_i q_{ij}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} u_j,$$

where σ is the elasticity of substitution, q_{ij} denotes the consumption of goods from location i, and u_j is the local amenity in location j.

The goods are produced in an Armington setting, where each location produces a distinct variety (with the same index as the producing location). The production of each variety

follows

$$q_i = A_i L_i,$$

where A_i and L_i are productivity and labor in location i, respectively. Each agent supplies one unit of labor inelastically and obtains wage w_i. Assuming perfect competition, we have

$$w_i = p_i A_i, \tag{7}$$

where p_i is the price of variety i in its home market. We denote by τ_{ij} the iceberg cost for goods produced in i and sold to location j. We assume the iceberg trade costs satisfy: $\tau_{ij} \geq 1$ and $\tau_{ij}\tau_{jk} \geq \tau_{ik}$. Thus, the price of variety i in j follows

$$p_{ij} = p_i \tau_{ij}. \tag{8}$$

The budget constraint of the representative agent in location i can thus be written as

$$\sum_j p_{ij} q_{ij} = w_i. \tag{9}$$

The representative agent's optimization, maximizing U_j subject to (9), leads to the gravity equation

$$X_{ij} = \tau_{ij}^{1-\sigma} p_i^{1-\sigma} P_j^{\sigma-1} Y_j, \tag{10}$$

where X_{ij} denotes the value of location j's imports from location i; $P_j \equiv \left(\sum_i p_{ij}^{1-\sigma}\right)^{\frac{1}{1-\sigma}}$ is the price index for location j and $Y_j = w_j L_j$ represents total income in the locality. The representative agent's welfare, denoted by W_j, follows

$$W_j = \frac{w_j}{P_j} u_j. \tag{11}$$

The goods market clearing condition is

$$Y_i = \sum_j X_{ij}. \tag{12}$$

Balanced trade gives

$$Y_i = \sum_j X_{ji}. \tag{13}$$

We allow for productivity or congestion externalities as in Allen and Arkolakis (2014),

$$A_i = \bar{A}_i L_i^{\alpha}, \tag{14}$$

$$u_i = \bar{u}_i L_i^{\beta}, \tag{15}$$

where $\alpha > 0$ and $\beta < 0$ govern the intensity of productivity and amenity spillovers, respectively. $\bar{A}_i$ and $\bar{u}_i$ are location-specific constant terms. Equation (11) then becomes

$$W_j = \frac{w_j}{P_j}\bar{u}_j L_j^{\beta}. \tag{16}$$

Free labor mobility implies welfare equalization: $W = W_i$. Then, the equilibrium conditions (12) and (13) become

$$w_i^{\sigma} L_i^{1-\alpha(\sigma-1)} = W^{1-\sigma} \sum_j \tau_{ij}^{1-\sigma} \bar{A}_i^{\sigma-1} \bar{u}_j^{\sigma-1} w_j^{\sigma} L_j^{1+\beta(\sigma-1)} \tag{17}$$

$$w_i^{1-\sigma} L_i^{\beta(1-\sigma)} = W^{1-\sigma} \sum_j \tau_{ji}^{1-\sigma} \bar{A}_j^{\sigma-1} \bar{u}_i^{\sigma-1} w_j^{1-\sigma} L_j^{\alpha(\sigma-1)} \tag{18}$$

Allen and Arkolakis (2014) prove that with free labor mobility, if $\alpha + \beta < 0$, $-1 < \beta < 0 < \alpha < 1$, there exists a unique regular equilibrium.[24]

4.2 Optimal Route Selection

Following Allen and Arkolakis (2019), we endogenize trade costs through traders' optimal route selection. Consider a transportation network connecting all the N locations. Some location pairs are directly connected while others are indirectly connected through other locations. We use an infrastructure matrix $T = [t_{ij}]$ to characterize the connected economy, where t_{ij} is the iceberg trade cost from city i to city j. $t_{ij} \in [1, \infty)$ if the two cities are directly connected, i.e., (i, j) is a city-pair link, and $t_{ij} = \infty$ otherwise.

Consider a continuum of traders $\nu \in [0, 1]$ who ship goods from i to j. They can either choose the direct link from i to j (if this direct link exists) or an indirect link through other locations that connect i and j. Denote $\Re_{ij}$ the set of all possible paths starting from i and ending at j, and $r \in \Re_{ij}$ represents one possible path. r_ι indicates the ι-th city along the path r:

$$r = \{r_0, r_1, \cdots, p_\iota, \cdots p_{K-1}, p_K\}$$

where K is the total length of path r, $r_0 = i$ and $r_K = j$. Any pair of $r_{\iota-1}$ *and* r_ι are directly connected for $\iota \in [1, K]$. The path-specific trade cost is

$$\tilde{\tau}_{ij}(r) = \prod_{\iota=1}^{K} t_{r_{\iota-1} p_\iota}. \tag{19}$$

In addition to the instantaneous trade costs, each trader would also face a path-specific idiosyncratic trade cost shock, $\varepsilon_{ij}(r, \nu)$, which is drawn from a *Weilbull* distribution:

$$Pr\left[\varepsilon_{ij}(r, \nu) \leq \varepsilon\right] = 1 - exp\left(-\varepsilon^{\theta}\right), \tag{20}$$

[24] An equilibrium is "regular" if $\forall i, L(i) > 0$.

where θ captures the heterogeneity of traders' preferences over routes. A higher θ implies less dispersed preferences and, hence, more agreements on the optimal route. Therefore, each trader's total trade cost from i to j along path r is $\tilde{\tau}_{ij}(r)\,\varepsilon_{ij}(r,\nu)$. The optimal route for trader ν solves

$$\tau_{ij}(\nu) = \min_{r} \tilde{\tau}_{ij}(r)\,\varepsilon_{ij}(r,\nu). \tag{21}$$

The optimized $\tau_{ij}(\nu)$ is trader ν's total trade cost from i to j. Applying the above equation to all traders, we obtain the average trade cost from i to j, $E_{\nu}[\tau_{ij}(\nu)]$, which is τ_{ij} in the economic geography model in Section 4.1,

$$\tau_{ij} = \left(c^{-\theta} \sum_{K=0}^{\infty} \sum_{r \in \Re_{ij,K}} \prod_{\iota=1}^{K} a_{r_{\iota-1} r_{\iota}} \right)^{-\frac{1}{\theta}} = \left(c^{-\theta} \sum_{K=0}^{\infty} \mathbf{A}_{ij}^{K} \right)^{-\frac{1}{\theta}} = \left(c^{-\theta} b_{ij} \right)^{-\frac{1}{\theta}}, \tag{22}$$

where $c \equiv \Gamma\left((\theta+1)/\theta\right)$, $\Re_{ijK}$ indicates the set of paths from i to j with length K, $a_{ij} \equiv t_{ij}^{-\theta}$, $\mathbf{A} \equiv [a_{ij}]$, $\mathbf{A}_{ij}^{K}$ is the ij-th element of $\mathbf{A}^{K}$, and $b_{ij} \equiv \sum_{K=0}^{\infty} \mathbf{A}_{ij}^{K}$.[25]

For now, let us take t_{kl} as exogenous parameters and examine their effects on trade costs. According to Allen and Arkolakis (2019), the elasticity of trade cost between any location pair, τ_{ij}, with respect to t_{kl} is[26]

$$\frac{d \ln \tau_{ij}}{d \ln t_{kl}} = \left(\frac{\tau_{ij}}{\tau_{ik} t_{kl} \tau_{lj}} \right)^{\theta}. \tag{23}$$

Allen and Arkolakis (2019) also prove that the elasticity can be interpreted as the probability to use link kl when a trader considers moving goods from i to j. Denote by Ξ_{kl} the total trade flow through link kl. We have

$$\Xi_{kl} = \sum_{i} \sum_{j} X_{ij} \left(\frac{\tau_{ij}}{\tau_{ik} t_{kl} \tau_{lj}} \right)^{\theta}. \tag{24}$$

This route selection framework embeds the economic geography model as follows. The optimal route selection endogenizes the determination of the expected trade cost τ_{ij} in (22), which serves as the direct input for the economic geography model to solve for trade flows X_{ij}. The latter can be used in the optimal route selection to recover the total trade flow through link kl, Ξ_{kl}, from (24).

4.3 Traffic Flow and Road Congestion

We assume that traffic flow on a (direct) link kl is determined by trade flows in the following way:

$$\ln(flow_{kl}) = \zeta + \lambda \ln(\Xi_{kl}), \tag{25}$$

[25] Here, $\mathbf{B} \equiv \sum_{K} \mathbf{A}^{K}$ if $\sum_{j} a_{ij} = \sum_{j} t_{ij}^{-\theta} < 1$ for $\forall i$.

[26] We use ij to denote any city pair, and kl to denote a directly-connected city pair. And for convenience, later we may call $\tilde{\tau}_{ij}$ as direct trade cost from k to l, and call τ_{ij} as expected trade cost from i to j.

where $\lambda \in (0, 1]$. Following Allen and Arkolakis (2019), we will let $\lambda = 1$ in the benchmark estimation and set λ to 0.5 as a robustness check.

$flow_{kl}$ determines $hours_{kl}$ by the congestion equation (2). We close the model by assuming t_{kl}, the trade cost on a direct link kl, to be determined by $hours_{kl}$:

$$\ln(t_{kl}) = \kappa \ln(hours_{kl}), \tag{26}$$

where $\kappa > 0$ and $hours_{klt}$ denotes the travel duration from k to l.

The equilibrium is defined as follows.

Definition 6 (Equilibrium)**.** Consider an economy associated with exogenous city-specific amenity $\bar{u}_i$ and productivity $\bar{A}_i$, and link-specific road capacity cap_{kl}. An equilibrium of the economy consists of city-specific w_i, L_i, p_i, P_i, Y_i, city-pair-specific τ_{ij}, X_{ij} and link-specific $hours_{kl}$, t_{kl}, Ξ_{kl}, $flow_{kl}$ and $traf_{kl}$ such that:

(i) Given $hours_{kl}$ and, hence, t_{kl} by (26), τ_{ij} is determined by the optimal route choices as in (22);

(ii) The goods market clearing condition (17) and balanced trade (18) pin down w_i and L_i,[27] which yield p_i, P_i, X_{ij} by the gravity equation (10) and Y_i by (12);

(iii) τ_{ij} and X_{ij} determine link-specific trade flow, Ξ_{kl}, through (24), which implies link-specific traffic flow, $flow_{kl}$, by (25);

(iv) $hours_{kl}$ and $flow_{kl}$ pin down $traf_{kl}$ by (1), which leads to $hours_{kl}^{new}$;

(v) $hours_{kl} = hours_{kl}^{new}$.

5 Structural Estimation

In this section, we structurally estimate the model. Following Allen and Arkolakis (2019), we set two pre-determined parameters: $\sigma = 9$ and $\lambda = 1$. As robustness checks, we will set $\lambda = 0.5$. The structural estimation is independent of α and β; for the welfare analysis we use $\beta = -0.3$ and $\alpha = 0.1$ as in Allen and Arkolakis (2019). The structural estimation also does not rely on exogenous city-specific amenity $\bar{u}_i$ and productivity $\bar{A}_i$, which will be calibrated in our welfare analysis, because we solve the economic geographic model by rationalizing city-level GDP data. This leaves three parameters to be structurally estimated: the shape parameter for traders' preference heterogeneity, θ; the cost parameter of $hours_{kl}$ in (26), κ; and the constant term in the traffic flow equation (25), ζ.

[27] A normalization like $\sum_i Y_i = 1$ or $P_i = 1$ for an i is necessary to pin down the equilibrium W.

Denote $\phi \equiv \{\kappa, \theta, \zeta\}$. Given ϕ, we simulate the average truck numbers (relative to the aggregate numbers) and driving hours for each direct link, denoted by $traf_{kl}^{sim}$ and $hours_{kl}^{sim}$, respectively.[28] We use nonlinear least square (NLS) to estimate ϕ, i.e., minimizing the distance between $traf_{kl}^{sim}$, $hours_{kl}^{sim}$ and their observed counterparts in the data, denoted by $traf_{kl}^{data}$, $hours_{kl}^{data}$. Note that, by construction, $\ln(hours_{kl}^{sim})$ are always identical to $\ln(hours_{kl}^{data})$ for uncongested links.

$$\hat{\phi} = \underset{\phi}{argmin} \sum_{kl} \left(\left(\ln\left(traf_{kl}^{sim}\right) - \ln\left(traf_{kl}^{data}\right) \right)^2 + \left(\ln\left(hours_{kl}^{sim}\right) - \ln\left(hours_{kl}^{data}\right) \right)^2 \right). \tag{27}$$

The estimation strategy is almost the same as Allen and Arkolakis (2019), but there is one important difference. Allen and Arkolakis (2019) do not include their data on hours from Google Maps in their NLS estimation. Our hours, instead, are from the real-time GPS data of actual trips, and we include this variable in the NLS.

θ and κ affect both $traf_{kl}^{sim}$ and $hours_{kl}^{sim}$. By construction, $traf_{kl}^{sim}$ is a relative measure and, hence, it is hard to recover ζ, the constant term in the traffic flow equation (25). However, ζ has a first-order effect on $hours_{kl}^{sim}$, which is not a relative measure. Therefore, the identification of ζ comes from $hours_{kl}^{sim}$.

5.1 Estimation Results

The estimation results are reported in Figure 3 and Table 4. Figure 3 shows that the estimated model fits the data well. $\hat{\phi}$ denotes the estimated ϕ. The correlation coefficient between $\ln(traf_{kl}^{sim(\hat{\phi})})$ and $\ln(traf_{kl}^{data})$ is 0.85, which is higher than the coefficient of 0.56 in Allen and Arkolakis (2019). $\ln(hours_{kl}^{sim(\hat{\phi})})$ fits $\ln(hours_{kl}^{data})$ almost perfectly, with a correlation coefficient of 0.98.

[28] The algorithm of solving $traf_{kl}^{sim}$ and $hours_{kl}^{sim}$ in the Section A.4 in the appendix. Allen and Arkolakis (2019) iterates on Ξ_{kl}. We iterate on $hours_{kl}$, because our data on $hours_{kl}$ can serve as a reasonable initial guess.

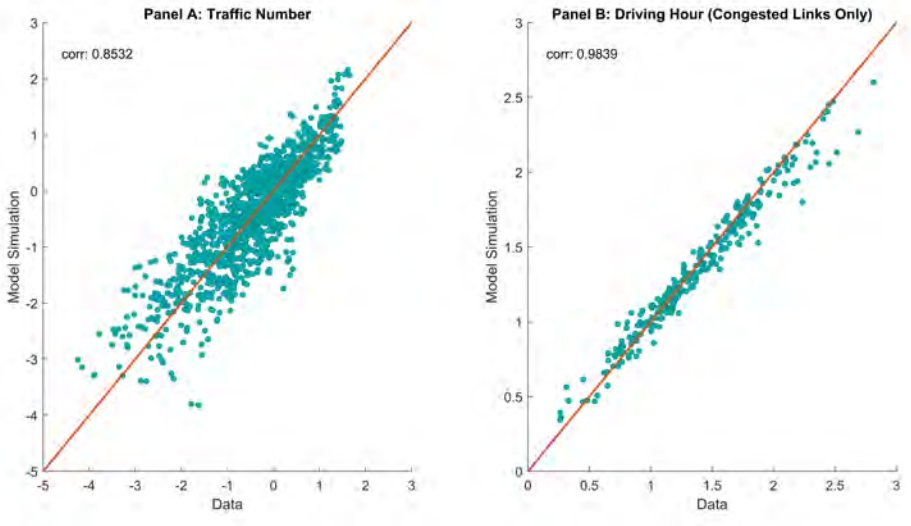


Figure 3: Model Fit for China

Note: Panel A and B plot $\ln(traf_{kl})$ and $\ln(hours_{kl})$ in the data and simulated from the estimated model, respectively. Panel B only shows congested links as the simulated $\ln(hours_{kl})$ are always identical to that in the data for uncongested links by construction.

For comparison, we also estimate the model using data from highways in China and England, respectively. One can see from Panel A of Figure 4 that the fit for $\ln(traf_{kl})$ is still good for China's highways. The correlation coefficient is 0.82. The fit just slightly decreases for England, with a correlation coefficient of 0.77 (Panel C). Both fits are overall comparable to that in Allen and Arkolakis (2019). For $\ln(hours_{kl})$, the match between the model and data is, again, almost perfect (Panel B and D).

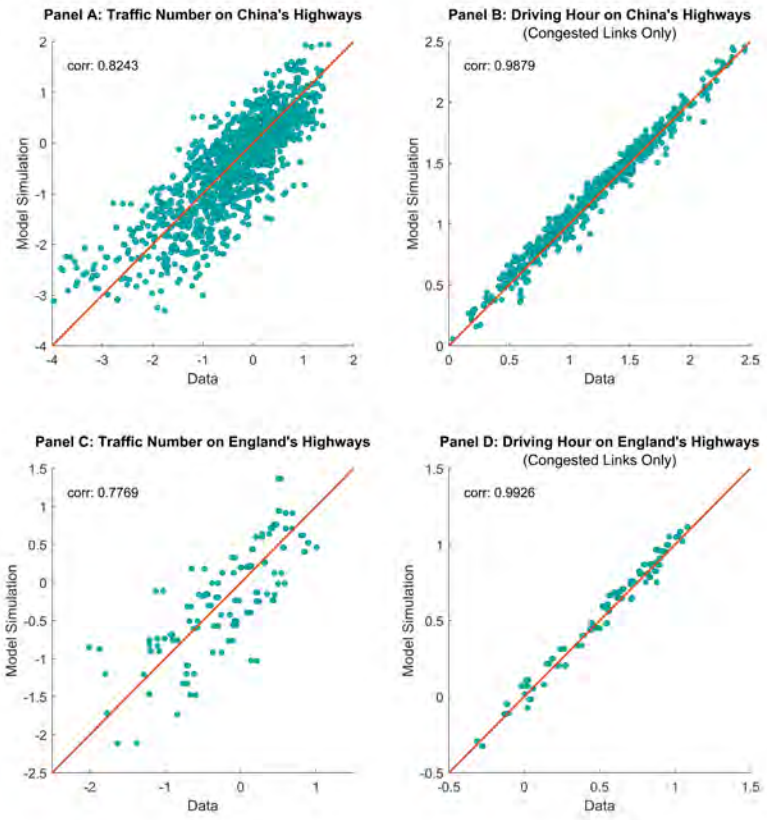


Figure 4: Model Fit under the Comparable Sample of China and England

Note: Panel A and C plot $\ln(traf_{kl})$ in the data and stimulated from the estimated model for highways in China and England, respectively. Panel B and D plot $\ln(hours_{kl})$ with the same structure. As in Figure 3, Panel B and D show congested links only.

The estimated parameter values are reported in the first three columns of Table 4. The estimated θ for China's NPCC roads is smaller than that for China's highways, which implies that drivers on highways behave more homogenously. This difference increases when comparing with England's highways, suggesting that more consensus is reached among drivers in England. In terms of the elasticity of direct trade costs to driving hours, κ, the result shows that the sensitivity to driving hours is slightly higher in China than in England. We also find that ζ is the largest in Column (1) and the smallest in Column (3). This suggests that, to transform the same amount of goods, more trucks are required on China's NPCC roads, but much fewer are needed on England's highways. We also report the estimates under $\lambda = 0.5$ in Column (4) to (6), and the results are overall robust to λ.

Table 4: Estimation Results

	(1)	(2)	(3)	(4)	(5)	(6)
	Benchmark: $\lambda = 1$			Robustness: $\lambda = 0.5$		
	China	China's Highways	England's Highways	China	China's Highways	England's Highways
θ	15.9295	20.7204	25.2605	16.4507	33.6462	97.4840
	(0.9321)	(0.9632)	(1.4961)	(0.0535)	(1.7124)	(0.5935)
κ	0.0418	0.0334	0.0302	0.0124	0.0182	0.0077
	(0.0023)	(0.0016)	(0.0014)	(0.0001)	(0.0009)	(0.0000)
ζ	-6.7952	-9.8581	-13.6211	-3.4333	-7.2145	-14.6492
	(0.1539)	(0.5132)	(1.5946)	(0.2043)	(0.6565)	(0.6271)

Note: Standard errors are reported in parentheses.

5.2 Indirect Inference

We further assess the quantitative performance of the estimated model by comparing the nontargeted moments predicted by the model and in the data. In particular, the structural estimation does not target explicitly the conditional relationships revealed by the empirical tests on the fundamental law of road congestion in Section 3.2.1. Therefore, we use indirect inference to investigate whether the model is consistent with these conditional correlations.

Specifically, we use traffic flow or vehicle-miles-travelled simulated by our model to replace the left-hand-side variable of the regression (6) and then run the same IV regression as in Section 3.2.1. The results are reported in Table 5.[29] The main conditional relationships are qualitatively similar. The estimated elasticity for congested links is above one, while that for uncongested links is below one. Their difference is also substantial and statistically significant, though the difference in the simulated data is quantitatively smaller than its empirical counterpart.

[29] A robustness check using the alternative IV with buffering distance of 50km is reported in Table 14.

Table 5: Indirect Inference with IV, Second Stage

Dep: VMT $(\ln)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Capacity $(\ln, lag)$	0.710***	1.290***	0.603***
	(0.138)	(0.330)	(0.175)
Speed limit $(\ln, lag)$	1.951***	0.579	2.242***
	(0.167)	(0.516)	(0.186)
Night light $(\ln)$	0.199*	1.580***	0.082
	(0.108)	(0.337)	(0.128)
Population in 1964 $(\ln)$	0.754***	0.523***	0.815***
	(0.055)	(0.100)	(0.071)
Urban land share in 1980	28.960***	6.362	30.984***
	(3.166)	(7.397)	(3.708)
Ruggedness	0.021	0.030	0.033
	(0.017)	(0.025)	(0.022)
Number of county pairs	0.000	-0.001	0.002
	(0.001)	(0.001)	(0.001)
N	1464	307	1157
R^2	0.728	0.779	0.705
p value for H_0		0.035	

Note: Robust standard errors in parentheses * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. The IV is the total length of historical road in a buffer of 10 kilometers extended from the city pair (excluding the length within the city pair). H_0: the coefficients of Capacity $(\ln, lag)$ in congested and uncongested groups are the same.

6 Welfare Implications

6.1 Elasticity and Welfare Gain

In this section, we use the estimated model to compute the welfare gain of expanding capacity of individual links. Following Allen and Arkolakis (2019), we derive an explicit formula for the elasticity of welfare with respect to link capacity:

$$\frac{d\ln W}{d\ln lane_{mn}} = \sum_{ij}\sum_{kl} \frac{d\ln W}{d\ln \tau_{ij}} \frac{d\ln \tau_{ij}}{d\ln t_{kl}} \frac{d\ln t_{kl}}{d\ln lane_{mn}}, \tag{28}$$

where $lane_{mn}$ denotes the average number of lanes on link (m,n) and we use variations in the number of lanes to represent variations in the capacity of the link.[30] Each of the three parts on the right hand side of (28) can be written as a function of parameters and observables. We present the formulas below but leave the derivations to the appendix.

First, the elasticity of aggregate welfare with respect to trade cost follows

$$\frac{d\ln W}{d\ln \tau_{kl}} = \begin{cases} -\frac{X_{kl}}{\sum_i Y_i} \times (1+\varsigma_k+\varpi_l) & if\ l \neq s \\ -\frac{X_{kl}}{\sum_i Y_i} \times (1+\varsigma_k) & if\ l = s, \end{cases}$$

where s refers to the city for which w_s is normalized to one. ς_k and ϖ_l are defined in Appendix A.6.1. In the special case without externalities (i.e., $\alpha = 0$ and $\beta = 0$) we obtain $\varsigma_k = 0$, $\varpi_l = 0$ and lowering trade cost on a link always yields aggregate welfare gains. The aggregate welfare elasticity is then exactly equal to the share of trade flows on the link in aggregate output, which coincides with the finding from Atkeson and Burstein (2010) and Lai et al. (2020) that the global gain from the reduction of trade cost can be captured by the share of trade flow among global GDP.

The second term, $d\ln \tau_{ij}/d\ln t_{kl}$, can be written as

$$\frac{d\ln \tau_{ij}}{d\ln t_{kl}} = \left(\frac{\tau_{ij}}{\tau_{ik}t_{kl}\tau_{lj}}c\right)^{\theta}, \tag{29}$$

where $c \equiv \Gamma\left((\theta+1)/\theta\right)$. The elasticity of τ_{ij} with respect to τ_{kl} is proportional to the ratio of the trade cost between city i and j to the trade cost on the route connecting city i and j through link (k,l). The elasticity, which is non-negative and smaller than 1, decreases in t_{kl}, implying a smaller effect of expanding the capacity of a link with higher t_{kl}. Meanwhile, we can intuitively expect that as θ increases (traders' heterogeneity decreases), traders' selections would be closer to the shortest path between i and j, and thus this elasticity (or equivalently, the possibility that one trader moving from i to j uses link kl) is more likely to increase if kl is quite close to the shortest path from i to j, but is more likely to decrease if kl is far away from such shortest path.

Third, putting (1), (25) and (26) into the congestion equation (2), and differentiating both sides w.r.t. $\ln lane_{mn}$, we obtain

$$\frac{d\ln t_{kl}}{d\ln lane_{mn}} = \frac{J_{kl}}{1-J_{kl}}\kappa\left(\lambda\sum_{k'l'}\frac{d\ln \Xi_{kl}}{d\ln t_{k'l'}}\frac{d\ln t_{k'l'}}{d\ln lane_{mn}} - \mathbf{1}\{kl=mn\}\right), \tag{30}$$

where $J_{kl} \equiv \delta_{1kl}\times\delta_{2kl}\times\left(\left(exp(\zeta)\times\Xi_{kl}^{\lambda}\times hours_{kl}\right)/\left(dist_{kl}\times lane_{kl}\right)\right)^{\delta_{2kl}}\times(dist_{kl}/hours_{kl})$, $J_{kl}/(1-J_{kl})$ is non-negative, and the specific formula of $d\ln \Xi_{kl}/d\ln t_{k'l'}$ is presented in Section A.6.4. An increase in lane capacity in a congested link (m,n) can directly decrease

[30] Similar to kl, mn also represents a directly-connected city pair.

$hours_{mn}$ and t_{mn}, which is the negative direct effect, and then affect trade costs and trade flows. Such effects will, in turn, change traffic flows and, hence, t_{kl} on each link. Trade costs and flows will be further affected until the system reaches a fixed point; this iteration to the fixed point yields the indirect effects, which can be either positive or negative. In other words, we can obtain the elasticity of t with respect to $lane$ by solving the equation system (30).[31]

6.2 Cost Calculation Procedure

We calculate the cost of improving each link by considering both the road construction cost and the opportunity cost of land. The road construction cost is calculated based on the national average value of 7.593 million RMB per lane-mile from official statistic reports,[32] and scaled for each link in proportion to its slope:[33]

$$unit\ construction\ cost_{kl} \propto 1 + slope_{kl}. \tag{31}$$

Specifically, we first calculate the average slope for each link using the "Slope" tool in ArcGIS based on Digital Elevation Model (DEM) data from the SRTM Digital Elevation Database with a spatial resolution 90 meters.[34] The slope proxies the relative values of unit construction costs as in (31). We can then pin down the unit construction cost for one specific link by requiring the weighted average of unit construction costs across all links, with each link's area as the weight, to equal the national average construction cost of 7.593 million RMB per lane-mile. This allows us to recover the unit costs for each link.[35]

Our approach regarding the opportunity cost of land is based on land price information for each year between 2014 and 2018 from https://www.landchina.com/. We geocode

[31]For uncongested links, the elasticity is zero since an increase in lanes on (m, n) cannot decrease $hours_{mn}$ or t_{mn}.

[32]According to the *Statistical Report of the Development of Transportation Industry* from 2014 to 2018, the total investment on road in during these years is 9,253,838 million RMB, and the increments of roads are 304,668 miles. We assume on average the roads are equipped with 4 lanes (2 lane for each direction), and then the total increment of general national and provincial roads is 1,218,672 lane-miles. We finally reach the average unit cost as 7.593 million RMB per lane-mile

[33]The relationship between construction costs and slope is similar to Faber (2014), but his equation also includes other determinants such as urban land, which is already at least partially captured through our data on land prices.

[34]The database is available at http://srtm.csi.cgiar.org/. We first calculate the slope for each raster, and then assign rasters to links by following the criteria that the distance is no larger than 1 km. The average slope for each link is the mean value of slopes across all rasters that are matched to that link.

[35]In practice, suppose we know the unit construction cost for one specific link. Since the slope determines the relative level of unit construction costs, all other unit construction costs can be computed as well. Then, by taking the weighted average of these unit construction costs and requiring the weighted average to equal 7.593, we can easily solve for the appropriate factor and then get the exact values for the construction costs of all links.

each land parcel as a point using GaoDe Map API, and assign these parcels to links according to the criteria that the distance is no greater than 500 meters. Then, for each segment, we can calculate the mean value of land prices by taking the average across all land parcels that are matched. By further taking average across all segments, we can obtain the average unit land price for each link.

By putting the above two parts together, we can obtain the total construction cost per lane-mile for each road. Following Allen and Arkolakis (2019), we assume an annual maintenance cost of 5%, under the assumption of a 20 year depreciation schedule.

6.3 Unequal Returns

We first compute the aggregate welfare gain of expanding the capacity of a link by adding one more lane to the link. The exercise is conducted as follows. By the equation for $d\ln W/d\ln lane_{kl}$ in (28), the aggregate welfare gain can be written as

$$\frac{dW}{dlane_{kl}} = \frac{d\ln W}{d\ln lane_{kl}} \times \frac{W}{lane_{kl}},$$

where W can be proxied by GDP. Figure 5 visualizes the aggregate welfare gains from expanding the capacity of each link in the structurally estimated model, where the parameter values follow the first column of Table 4. The basic statistics are summarized in Column (1) of Table 6.



Figure 5: Welfare Gains from One Additional Lane

We next discuss the cost side. The unit costs for improving each link are presented in Figure 6, where the left panel focuses on construction costs (million RMB per lane-mile) and the right panel corresponds to the opportunity costs (million RMB per square miles). It is consistent with our expectation that the construction cost related to topography is much higher in mountainous south-eastern and south-western areas, but the land prices are much higher in more developed areas like Shanghai, Shenzhen, and Beijing. We then compute the total cost of one more lane for each link. For the construction cost, multiplying the unit cost with the length of each link yields the total cost. For the opportunity cost of land, by assuming the average lane width of 3.5 meters, we can calculate the area (in terms of squared miles) corresponding to one additional lane for each link, and obtain the total opportunity cost by similarly multiplying the unit cost with the corresponding area. The basic statistics are summarized in Columns (2) and (3) of Table 6.

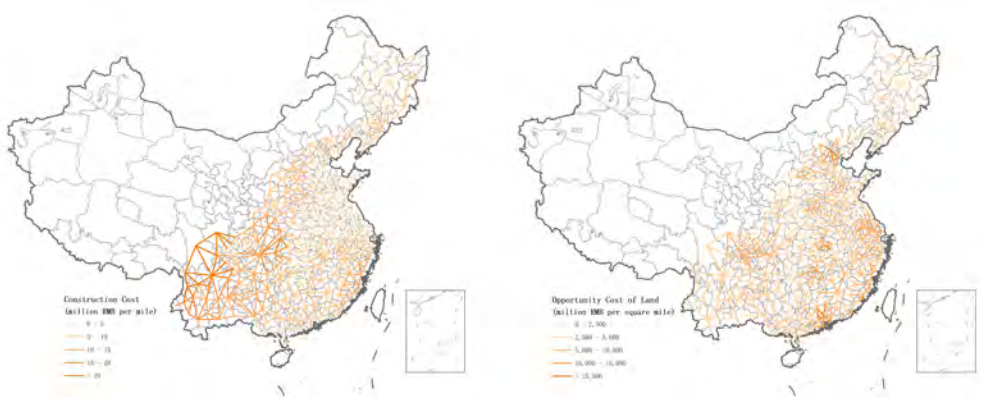


Figure 6: Unit Costs from One Additional Lane

Note: The left panel visualizes the spatial distribution of unit construction costs, and the right panel corresponds to the unit opportunity costs of land.

Table 6: Summary Statistics of Benefits and Costs from One Additional Lane

	(1)	(2)	(3)	(4)	(5)
	Annual Benefit	Total Construction Cost		Total Opportunity Cost of Land	
	Congested Links	Congested Links	All Links	Congested Links	All Links
mean	167.861	1123.462	1115.495	1325.411	1036.995
std	159.197	1446.31	1374.153	897.156	672.102
min	5.909	94.153	69.770	236.835	51.501
25%	82.568	315.256	302.182	720.098	593.781
50%	122.212	647.845	663.377	1004.419	872.845
75%	201.401	1288.229	1367.496	1655.667	1272.778
max	1193.299	10310.949	11327.696	4808.624	4808.624

Note: The unit is million RMB.

Based on the benefits and costs above, we calculate the return of road expansion following the formula

$$R_{kl} = \frac{B_{kl} - 5\% \times (C_{g,kl} + C_{l,kl})}{C_{g,kl} + C_{l,kl}}, \tag{32}$$

where $B_{kl} \equiv W \times d \ln W / d \ln lane_{kl}$ indicates the annual benefit from one additional lane, $C_{g,kl}$ and $C_{l,kl}$ represent topography-related construction cost as well as opportunity cost of land for one additional lane, respectively, and 5% is the annual maintenance cost. We summarize the return in Column (1) of Table 7. While only 20% of the links are congested in China, we find that for these congested ones the average return is around 5.2%. Meanwhile, we also find a large variation of returns, as illustrated in Figure 7. We report the top ten links in terms of the returns in Table 17. The opportunity cost of land is usually not directly

measured in the existing literature or approximated with urbanization.[36] Our results suggest that the opportunity cost of land is crucial for the quantification of welfare effects of roads in China. Not only does including the opportunity cost reduce the average return substantially (see Columns (1) and (2) in Table 7), it also affects the spatial distribution.[37]

Table 7: Gross Returns to One Additional Lane in China

	(1) Benchmark	(2) Excluding Land Cost	(3) $\lambda = 0.5$	(4) No Labor Mobility	(5) No Externality ($\alpha = 0$ and $\beta = 0$)
mean	0.052	0.498	0.041	0.064	0.056
std	0.104	1.117	0.100	0.162	0.135
min	-0.050	-0.049	-0.050	-0.050	-0.050
25%	-0.016	0.019	-0.030	-0.024	-0.022
50%	0.023	0.135	0.004	0.014	0.011
75%	0.095	0.526	0.076	0.103	0.087
max	0.789	9.814	0.754	1.528	1.239

Note: Only congested links are reported in this table, and the returns for uncongested links are zero.

The same counterfactual exercise is also conducted in the structurally estimated model with $\lambda = 0.5$, where the parameter values follow Column (4) in Table 4. The results are reported in the third column of Table 7. Externalities and labor mobility do not affect our structural estimation. However, they do matter for welfare implications.[38] We check the robustness of the above results by conducting the same counterfactual exercise in the model without externality and labor mobility, respectively. The results are reported in the fourth and fifth columns of Table 7. The differences are quantitatively small and the returns are robust to the alternative model specifications.

[36] See, e.g., Faber (2014), Alder (2019), and Allen and Arkolakis (2019).

[37] The spatial distribution of the returns when the opportunity cost of land is excluded is reported in Figure 20.

[38] See Donaldson and Hornbeck (2016) that assumes perfect mobility and Tombe and Zhu (2019) for limited mobility.

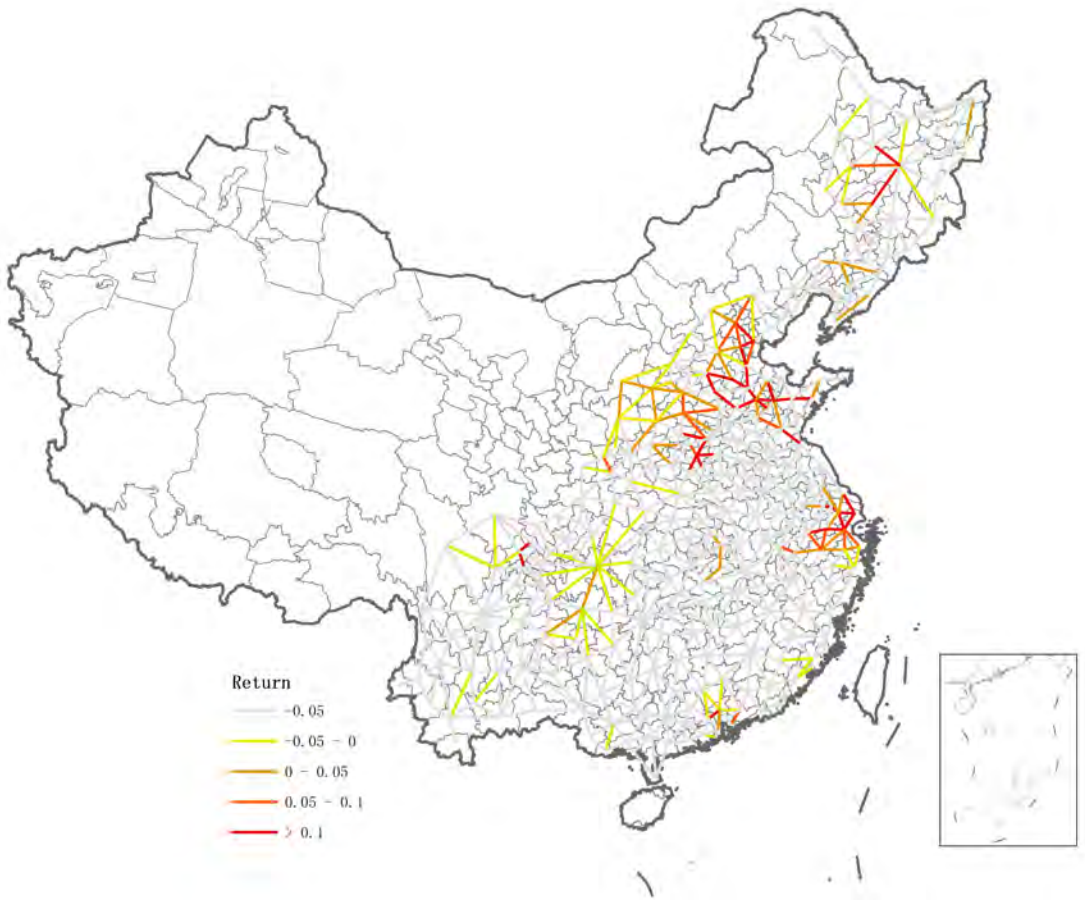


Figure 7: Spatial Distribution of Returns

6.4 Aggregate Returns between 2013 and 2017

Equipped with the above return from one additional lane, we are able to calculate the aggregate return by taking into account the actual (and uneven) expansion of each link during China's development. Specifically, we can use the lane information at the end of 2013 and 2017, and thus the actual expansion of each link during this period can be calculated. Then, we calculate the aggregate return as follows

$$R = \frac{\sum_{kl} B_{kl} \times \Delta lane_{kl} - 0.05 \times \sum_{kl}(C_{g,kl} + C_{l,kl}) \times \Delta lane_{kl}}{\sum_{kl}(C_{g,kl} + C_{l,kl}) \times \Delta lane_{kl}}, \tag{33}$$

where $B_{kl} \equiv W \times d\ln W/d\ln lane_{kl}$ indicates the annual benefit from one additional lane, $C_{g,kl}$ is the topography-related construction cost for one additional lane, $C_{l,kl}$ is the opportunity cost of land for one additional lane, and $\Delta lane_{kl}$ represents the lane increase for link kl between 2013 and 2017. This aggregated value is reported in Table 8. We find that the aggregate return is negative, mainly because around 80% of links are uncongested and

generate zero benefits. When we restrict our attention to the congested links, then the return is is 2.3% when considering the opportunity cost of and 10.5% when excluding it, which are slightly smaller than the mean value in Table 7, and is consistent with our finding that the return is negatively correlated with the observed road expansion (Figure 21).

Table 8: Aggregate Gross Returns between 2013-2017

	Aggregate Gross Return (including land cost)		Aggregate Gross Return (excluding land cost)	
	All Links	Congested Links	All Links	Congested Links
2018	-0.032	0.023	-0.016	0.105
2014	-0.036	0.010	-0.026	0.056
2014 + productivity growth	-0.030	0.033	-0.018	0.096

Note: For 2018 (2014), we use road infrastructure at the beginning of 2018 (2014), or equivalently at the end of 2017 (2013).

6.5 Productivity Growth and Returns

Our exercises reveal highly unequal returns to road investment. Furthermore, the aggregate returns to road investment are surprisingly low. An obvious caveat is that the static model tends to under-estimate the returns in a fast-growing economy like China. Expanding a congested road, albeit low returns under the current economic condition, might deliver much higher future returns if strong productivity growth can be anticipated. Productivity growth could even turn an uncongested road into a congested one and, therefore, justify some seemingly futile investments. While we are unable to forecast productivity growth, it is useful to redo all the welfare calculations in an earlier year by assuming perfect foresight of economic fundamentals in 2018. Specifically, consider an economy in which the city-to-city links follow China's road network in 2014 and the city productivity and amenity are set equal to those calibrated to city GDP and population in that year. In other words, the economy is calibrated to fit China's economic geography in 2014. The parameters, θ, κ and ζ, are the same as the structurally estimated values. We conduct two exercises. The first is to do the same welfare calculations for returns to road investment in the 2014 economy. The second is to evaluate the returns in the economy with the road network in 2014 but productivities and amenities in 2018.

In the first exercise, we calculate the returns to road investment as in the benchmark model by using the number of lanes in each link at the end of 2013, the earliest year when the information is available, GDP and population in each city in 2014. In the second exercise,

we back out productivity and amenity in each city using city GDP and population in 2018.[39] The same method can also back out city productivity in 2014. Figure 22 shows a significant productivity growth in most cities. The annualized productivity growth has an average of 6.08% across the cities. We then impose the 2018 productivities and amenities into a model with the 2014 link capacity and simulate GDP, population and traffic flows in the model. The simulated data are used to calculate the returns to road investment in the second counterfactual.

Figure 8 plots the returns in the two experiments. As expected, productivity growth increases the returns in the second counterfactual. The difference is quantitatively important. Table 8 shows that the aggregate return in the congested links is 3.3% in the second exercise, about three times higher than those in the first exercise. The difference of the aggregate returns is also quantitatively sizable. In other words, returns to road investment are indeed sensitive to economic growth. Nevertheless, the returns in the first exercise are strongly correlated with those in the second exercise. Replacing productivities and amenities in 2018 with those in 2014 does not change our main finding of highly unequal returns to road investment in China. The huge dispersion of the returns would not narrow unless productivity converges rapidly across cities, a pattern that has not yet appeared in the data (see Figure 22 for the highly persistent productivity disparity).

[39] $\bar{A}$ and $\bar{u}$ can be calibrated to match city GDP and population through the production function and welfare equalization.

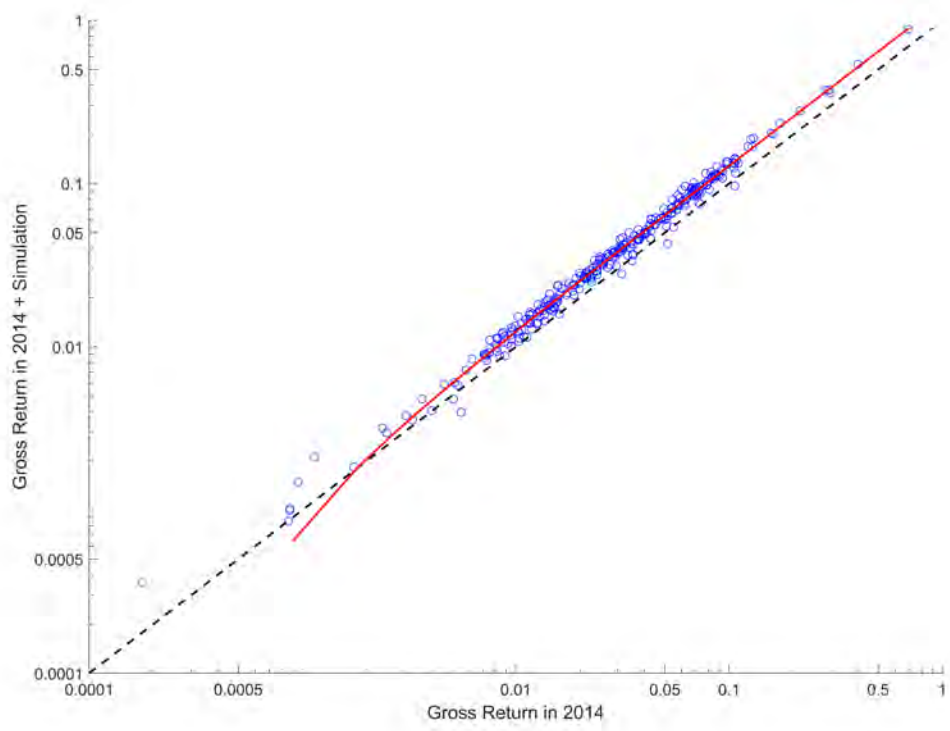


Figure 8: Comparison of Returns between Scenarios

Note: The black-dashed line is the 45 degree line and the red-solid line is the polynomial fitted line.

6.6 Comparison of Welfare Elasticities between China and England

Finally, we compare the welfare elasticity between China and England under the comparable sample in Table 9. The elasticity for England turns out to be much larger than that for China: The mean value of the welfare elasticity for England is almost ten times as large as that for China. This comparison not only indicates the huge contrast in the current stage between China and England, but also implies the potential level of welfare gain when China moves to the steady state like England (with very high road density and low road investment) in the future.

Table 9: Comparison of Welfare Elasticity between China and England

	China	England
mean	0.023	0.190
std	0.013	0.092
min	0.002	0.043
25%	0.013	0.132
50%	0.021	0.178
75%	0.030	0.236
max	0.087	0.617

Note: For England, we only have information for highways. To ensure the comparability between these two counties, we restrict our attention to highways for China. The table reports welfare elasticity multiplied by 10,000.

7 Conclusion

We quantify the returns to road infrastructure investments in China using a general equilibrium trade model with congestion and detailed data on traffic flows from GPS devices. The model captures the key features of the observed traffic patterns and we show that congestion is heterogeneous across links. We then use the estimated model to predict the returns to expanding road capacity in each link separately. The results suggest that the return to expanding the capacity of a link is low on average and dispersed across the network. We show that the unequal returns are hard to reconcile with the observed spatial distribution of productivity growth and likely a manifestation of misallocation of infrastructure investments.

Admittedly, model misspecification may also contribute to the dispersion of estimated returns to road investments. However, our identification of congested and uncongested links is model free. Expanding a uncongested link would always deliver zero benefits and negative returns unless China is undergoing a dramatic transformation of its economic geography that turns the uncongested links into congested ones and, yet, does not make the congested links more congested. The hypothetical economic geographic transformation implies a large-scale population flow from coastal to inland areas, which is exactly the opposite to what has been happening over the past four decades.

While we find it hard to justify the huge dispersion of returns to road investments from the perspective of maximizing aggregate productivity, there might be other compelling reasons that are missing in our analysis. For instance, we find that returns to investments in a city-to-city link positively correlate to the average GDP per capita in the city pair. Expanding a low-return link can serve as a policy instrument to reduce income inequality across cities.

Finally, the primary focus of our paper is on the allocation efficiency of investment within

road infrastructures. Yet, our estimates may also shed light on the allocation efficiency of investment between road infrastructures and other sectors. The low returns to road investments in China are actually consistent with the optimal allocation of investment between sectors. China is still holding large foreign reserves, most of which are US treasury bills with nearly zero returns. China's aggregate returns to capital also fell from about 15% in the mid 2000s to below 5% in recent years (see, e.g., Bai et al., 2016; Chen et al., 2019). The evidence we find is suggestive of severe misallocation in China's road investments. Yet, it should not be taken as evidence for overinvestment in road infrastructure. In contrast, although misallocation of road investments appears to be far less of a concern in England, the much larger welfare elasticities may imply very high returns to road investments and, therefore, underinvested road infrastructures.

Bibliography

Alder, Simon, "Chinese Roads in India: The Effect of Transport Infrastructure on Economic Development," 2019. working paper.

_ **and Illenin Kondo**, "Political Distortions and Infrastructure Networks in China: A Quantitative Spatial Equilibrium Analysis," 2019. working paper.

Allen, Treb and Costas Arkolakis, "Trade and the Topography of the Spatial Economy," *The Quarterly Journal of Economics*, 2014, *129* (3), 1085–1140.

_ **and** _ , "The Welfare Effects of Transportation Infrastructure Improvements," 2019. working paper.

_ **, Arkolakis Costas, and Yuta Takahashi**, "Universal Gravity," *Journal of Political Economy*, 2020, *128* (2).

_ **, David Atkin, Santiago Cantillo, and Carlos Hernandez**, "Trucks," 2020.

Asker, John, Allan Collard-Wexler, and Jan De Loecker, "Dynamic Inputs and Resource (Mis)allocation," *Journal of Political Economy*, 2014, *122* (5), 1013–1063.

Atkeson, Andrew and Ariel Tomas Burstein, "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy*, 2010, *118* (3), 433–484.

Bai, Chong-En Bai, Chang-Tai Hsieh, and Zheng Song, "The Long Shadow of A Fiscal Expansion," *Brookings Papers on Economic Activity*, 2016, *47* (2), 129–181.

Banerjee, Abhijit, Esther Duflo, and Nancy Qian, "On the Road: Access to Transportation Infrastructure and Economic Growth in China," *Journal of Development Economics*, 2020, *145* (14), 102442.

Baum-Snow, Nathaniel, "Did Highways Cause Suburbanization?," *The Quarterly Journal of Economics*, 2007, *122* (2), 775–805.

_ , J. Vernon Henderson, Matthew A. Turner, and Qinghua Zhang, "Does Investment in National Highways Help or Hurt Hinterland City Growth?," *Journal of Urban Economics*, 2020, *115*, 103124.

Bils, Mark, Peter J. Klenow, and Cian Ruane, "Misallocation or Mismeasurement?," 2020. National Bureau of Economic Research Working Paper Nr. 26711.

Chandra, Amitabh and Eric Thompson, "Does Public Infrastructure Affect Economic Activity? Evidence from The Rural Interstate Highway System," *Regional Science and Urban Economics*, 2000, *30* (4), 457–490.

Chen, Jingjing, Wei Chen, Ernest Liu, Jie Luo, and Zheng (Michael) Song, "The Economic Impact of COVID-19 in China: Evidence from City-to-City Truck Flows," 2020. working paper.

Chen, Wei and H. Allen Klaiber, "Does Road Expansion Induce Traffic? An Evaluation of Vehicle-Kilometers Traveled in China," *Journal of Environmental Economics and Management*, 2020.

_ , Xilu Chen, Chang-Tai Hsieh, and Zheng Song, "A Forensic Examination of China's National Accounts," *Brookings Papers on Economic Activity*, 2019, *Spring*, 77–127.

Donaldson, Dave, "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure," *American Economic Review*, 2018, *108* (4-5), 899–934.

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

Duranton, Gilles and Matthew A. Turner, "The Fundamental Law of Road Congestion: Evidence from US Cities," *American Economic Review*, 2011, *101* (6), 2616–2652.

_ and _ , "Urban Growth and Transportation," *Review of Economic Studies*, 2012, *79* (4), 1407–1440.

Faber, Benjamin, "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System," *Review of Economic Studies*, 2014, *81*, 1046–1070.

Fajgelbaum, Pablo D. and Edouard Schaal, "Optimal Transport Networks in Spatial Equilibrium," *Econometrica*, 2020, *88* (4), 1411–1452.

Fay, Marianne, Sungmin Han, Hyoung Il Lee, Massimo Mastruzzi, and Moonkyoung Cho, *Hitting the Trillion Mark: A Look at How Much Countries Are Spending on Infrastructure*, The World Bank, 2019.

Gurara, Daniel, Mr Vladimir Klyuev, MissNkunde Mwase, Andrea Presbitero, Xin Cindy Xu, and Mr Geoffrey J Bannister, *Trends and Challenges in Infrastructure Investment in Low-Income Developing Countries*, International Monetary Fund, 2017.

He, Guojun, Yang Xie, and Bing Zhang, "Expressways, GDP, and The Environment: The Case of China," *Journal of Development Economics*, 2020, *145*, 102485.

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

Hsu, Wen-Tai and Hongliang Zhang, "The Fundamental Law of Highway Congestion Revisited: Evidence from National Expressways in Japan," *Journal of Urban Economics*, 2014, *81*, 65–76.

Krishna, Pravin and Eva van Leemput, "Quantifying Sources of Internal Trade Barriers: Evidence From Indian Truck GPS Data," 2018.

Lai, Edwin L-C., Haichao Fan, and Han Steffan Qi, "Global Gains from Reduction in Trade Costs," *Economic Theory*, 2020, *70*, 313–345.

Qin, Yu, "China's Transport Infrastructure Investment: Past, Present, and Future," *Asian Economic Policy Review*, 2016, *11*, 199–217.

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

Small, Kenneth A. and Erik T. Verhoef, *The Economics of Urban Transportation*, Routledge, 2007.

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

Tombe, Trevor and Xiaodong Zhu, "Trade, Migration, and Productivity: A Quantitative Analysis of China," *American Economic Review*, 2019, *109* (5), 1843–1872.

Vickrey, William, "Optimization of Traffic and Facilities," *Journal of Transport Economics and Policy*, 1967, *1* (2), 123–136.

Àngel Garcia-López, Miquel, Ilias Pasidis, and Elisabet Viladecans-Marsal, "Congestion in Highways When Tolls and Railroads Matter: Evidence from European Cities," 2020. working paper.

A Appendix

A.1 Chinese Data, Additional Figures and Tables

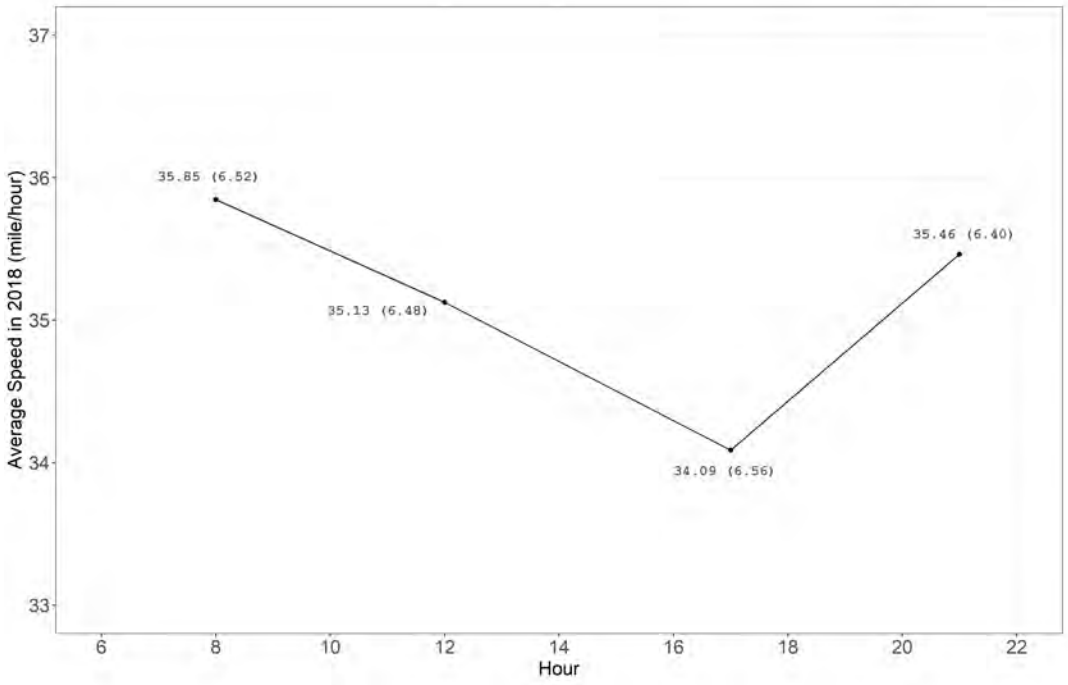


Figure 9: Temporal Speed Distribution

Note: The numbers (outside the parentheses) indicate the average speed and the number in the parentheses are standard deviations. On the horizontal axis, 8 indicates 8:00-9:00, 12 indicates 12:00-13:00, and so on for other hours.

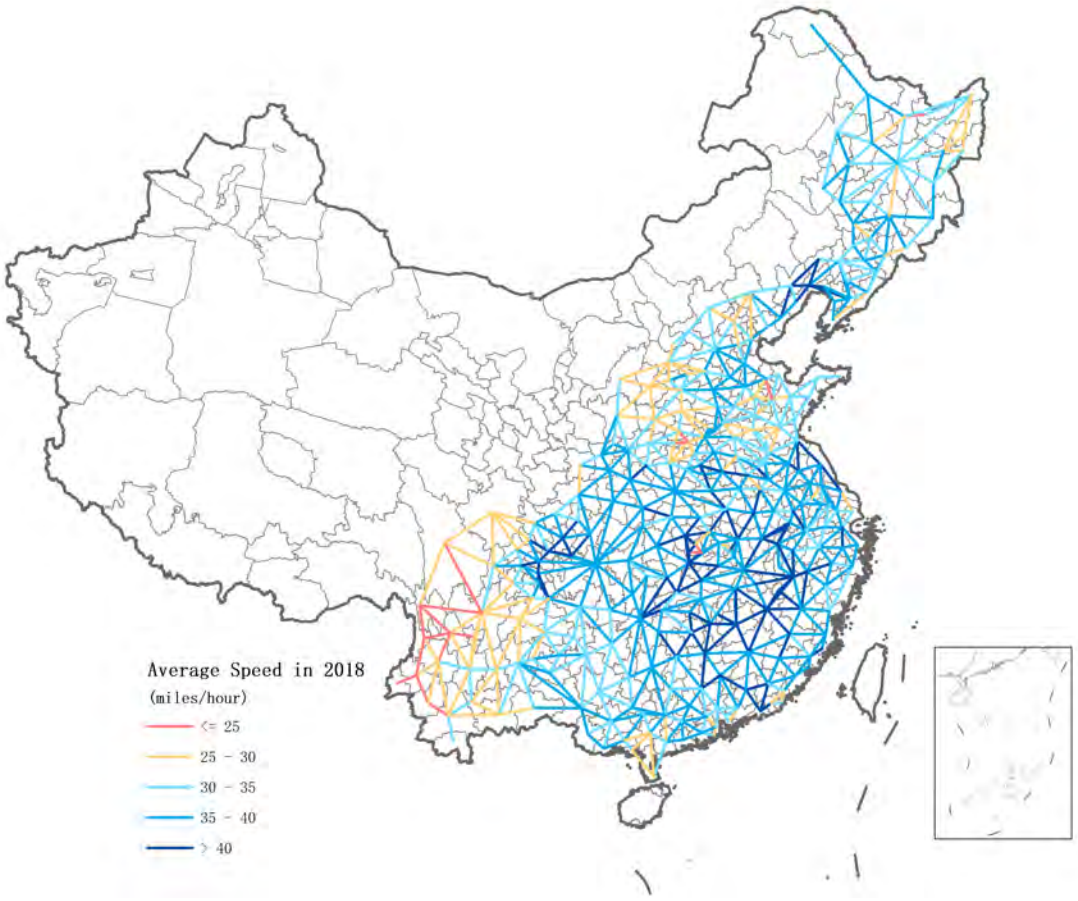


Figure 10: Spatial Speed Distribution

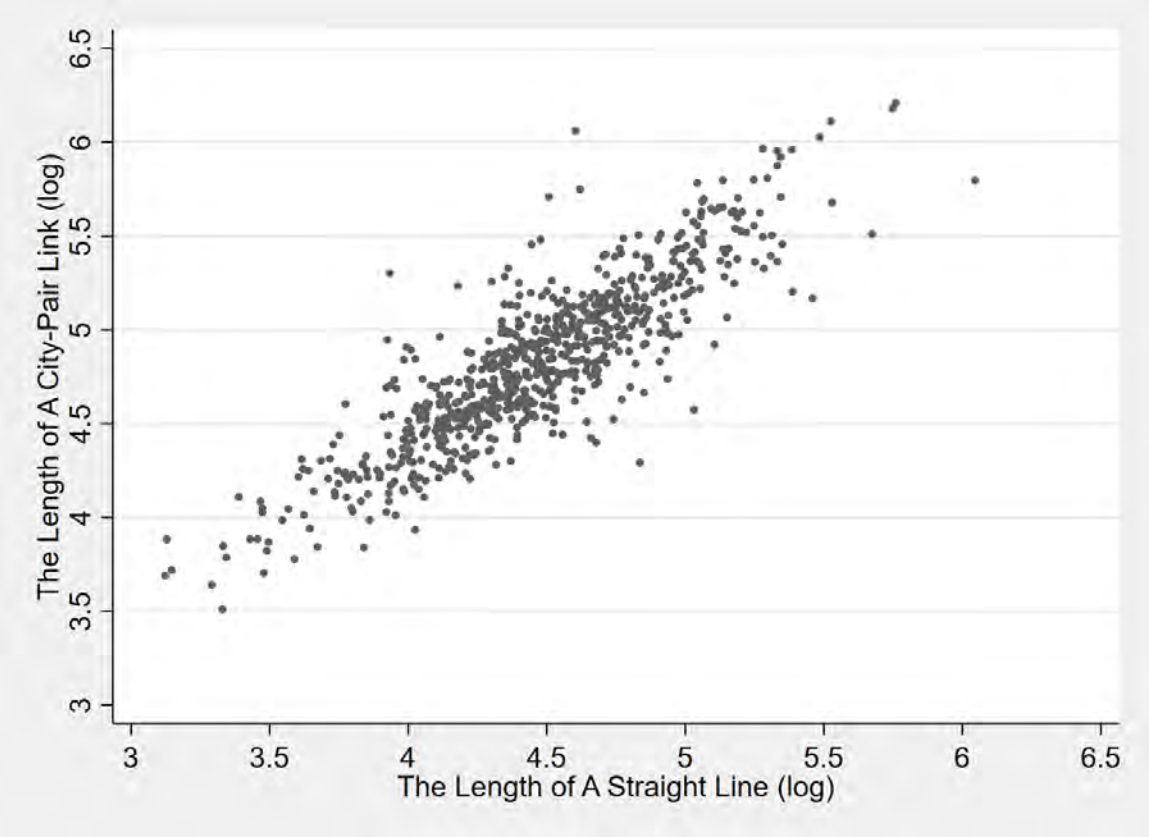


Figure 11: Comparison between Average Length and Straight-Line Length

Note: In the regression of average length on straight-line length (both in log), the coefficient is 0.897, with standard error of 0.014 and a R^2 of 0.759.

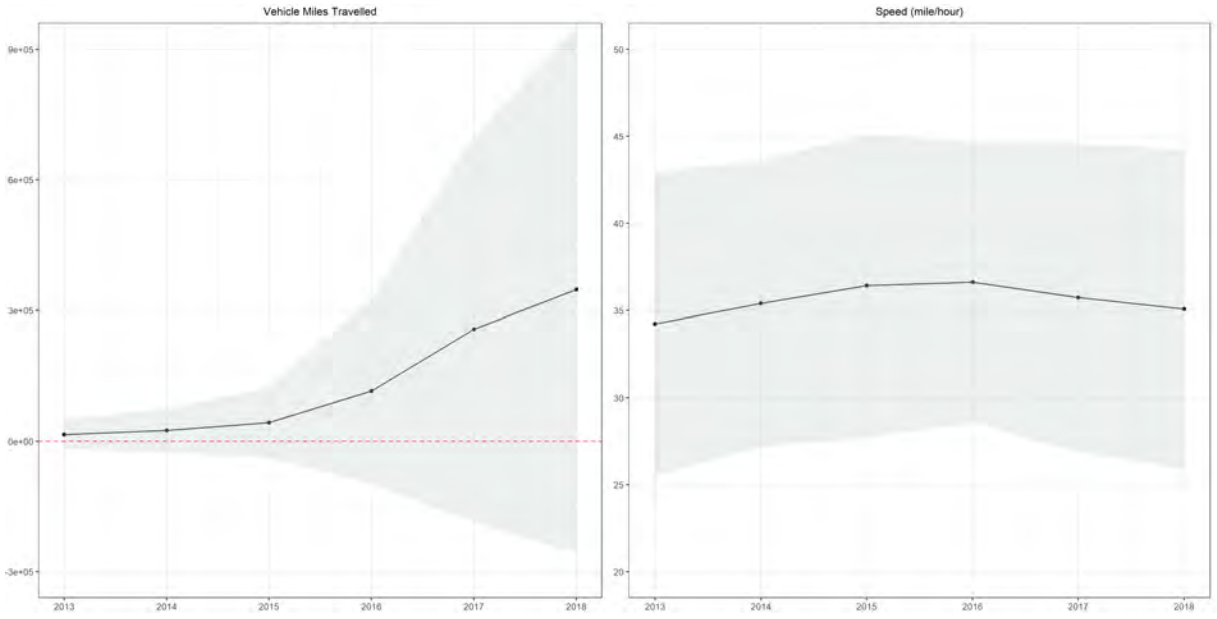


Figure 12: VKT v.s. Speed across Years

Note: The shade areas are 95% confidence intervals.

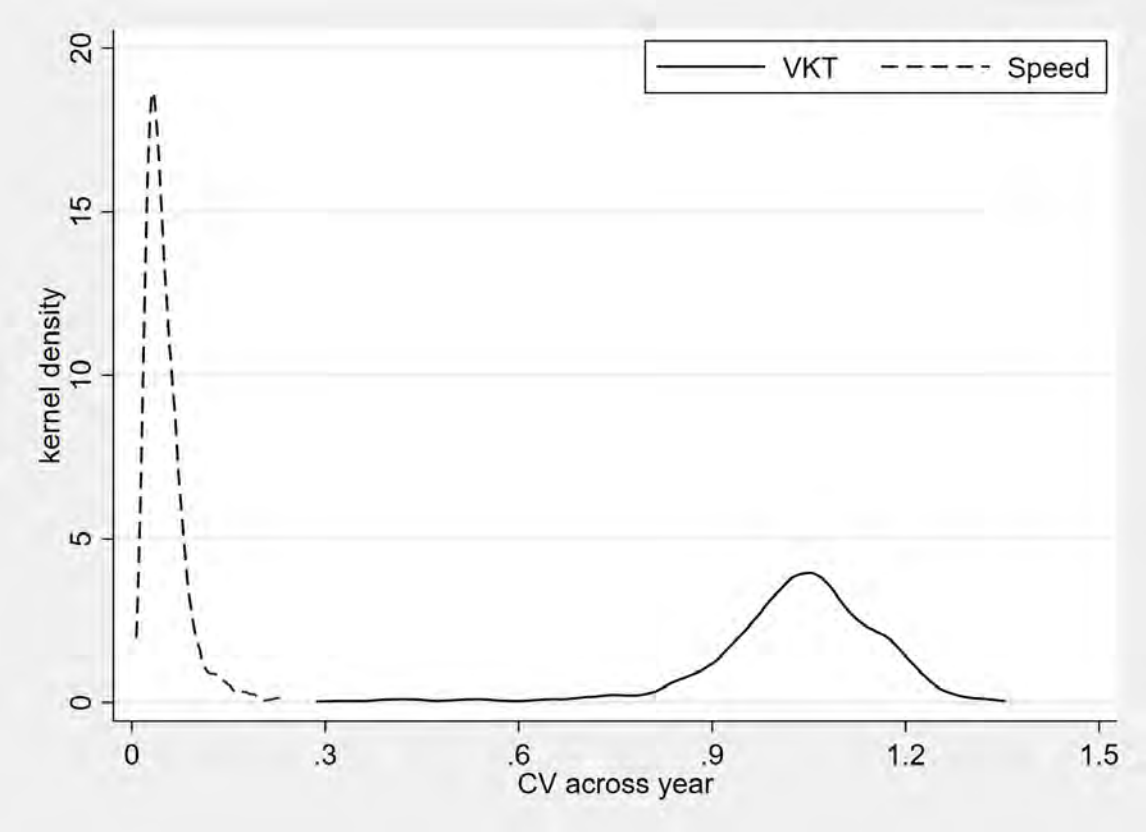


Figure 13: Coefficient of Variation of VKT and Speed across Years

Note: This figure plots the distribution of the CV (across years) of VKT and Speed of each citypair.

A.2 England Data

England has detailed data on traffic volumes and speed at thousands of measurement sides installed along its highway network, which we use to compare our results from the Chinese traffic data based on GPS devices on trucks. As for China, we complement the traffic data with information on the road network and regional GDP data.

A.2.1 City and Road Network

We define the national highway network in England as all roads that are maintained by *Highways England*, which is a national government company that is responsible for the motorways and major (trunk) roads in England. Figure 14 shows the extent of the network.

We define the administrative units in England following the Nomenclature of Territorial Units for Statistics (NUTS 2), which divides England into 33 regions. We restrict the sample to the 28 regions that are connected by the highways maintained by *Highways England* and have measurement sites with non-missing observations. Gross domestic product of each region is available from the Office for National Statistics.[40] The nodes of the infrastructure network correspond to the points on the network that are closest to the centroids of each region.

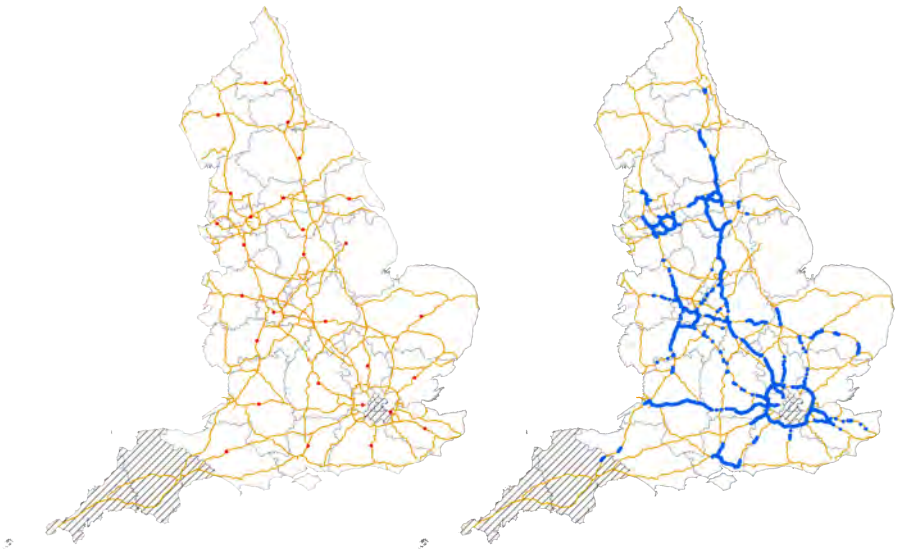


Figure 14: England's Roads, Regions, and Measurement Sites

Note: The non-shaded areas are the 28 regions in our sample, and the orange lines indicate the roads managed by *Highway England*. The red points on the left map represent the points on the roads that are nearest to the centroids of each region, and the blue points represent the measurement sites.

[40] See https://www.ons.gov.uk/economy/grossdomesticproductgdp/datasets/regionalgrossdomesticproductallnutslevelregions

A.2.2 Traffic Data from Measurement Sites

We use traffic data from Highway England's Motorway Incident Detection and Automatic Signalling (MIDAS), which monitors traffic using sensors installed in the road. The right map in Figure 14 shows the distribution of the measurement sites. The data contains information on traffic flow and speed for more than 5,000 measurement site during each 15-minute interval.[41] A highway link between two nodes can have many measurement sites and we average speed and density across all measurement sites on the same link in order to have a comparable data set as for China.

For the model estimation and counterfactual analyses, we aggregate the measurement sites to links, because each link in our network has multiple measurement sites. We use the average speed over all measurement sites on link kl within each 15-minute interval to measure $speed_{kl}$. We calculate the density at each measurement site by dividing traffic flow by speed, and adjust it using lane numbers.[42] Then, we use the average density across all measurement sites on link kl as $density_{kl}$.[43] Finally, to be consistent with our data management for China's data, we also follow equation (1) to transform $density_{kl}$ to $flow_{kl}$.

A.2.3 Comparison to Congestion in England

We now apply the same method to estimate congestion in England. The goal is to compare China with a mature economy and to make sense of its small share of congested links. Before preceding, let us first compare highways and their related economic conditions between the two economies. The reason for focusing on highways is that the MIDAS data we will use only covers highways in England.

Table 10 reports the summary statistics.[44] The first and third columns are for China and England. The middle column reports the statistics for three economically advanced regions in China that are more comparable to England. The first seven rows compare China and England

[41] For a sub-sample of the measurement sites, we also collected data on travel times from Google Maps. The travel speed implied by the Google data is consistent with the speed in the MIDAS data. The advantage of the MIDAS data, is that we also have the traffic flow, which allows us to analyse congestion.

[42] The traffic flow is measured as the number of vehicles per 15-minute interval. Hence, the unit of traffic density is $\frac{\text{number of vehicles/15-min}}{\text{km/15-min}} = \frac{\text{number of vehicles}}{\text{km}}$. As the unit of density is $\frac{\text{number of vehicles}}{\text{km} \times \text{number of lanes}}$, we need to divide the previous term by the number of lanes to get density.

[43] The road network is defined at the level of links and we abstract from heterogeneity across measurement sites on the same link by using the average for each link.

[44] The gridded GDP data are obtained from the *Global Dataset of Gridded Population and GDP (1980-2010 Estimations and 2020-2100 Scenarios)* for 2010, where the grid size is 0.5×0.5 degree and the unit is Billion $US\$2005/yr$ in PPP. The data are available at http://www.cger.nies.go.jp/gcp/population-and-gdp.html. For population, we do not refer to the same source because it is constructed for every ten years and thus there is no information for 2018. The population data for 2018 are obtained from *LandScan Datasets*, where the resolution is 1 km: https://landscan.ornl.gov.

on the capacity of highways (the first row) and their density (from the second to seventh rows). Not surprisingly, the statistics suggest a huge difference between China and England. In particular, despite the rapid expansion of China's highway system over the past two decades, the density of highways in China is still about half of that in England. Interestingly, the density of highway in the three economically advanced regions in China is very similar to that in England. The last three rows compare the economic conditions between the two economies. Again, the three economically advanced regions are more comparable to England, though the income gap is still sizable.

Table 10: Comparison of China and England's Highways

	China	China Yangtze River Delta, Pearl River Delta and Beijing-Tianjin-Hebei	England
Lane number	4.27	4.89	4.56
Highways' length density	2.70	4.01	5.09
Highways' area density	11.53	19.62	23.22
Average distance from grids (1km) to highways	24.54	11.84	10.46
Average distance from grids (10km) to highways	24.56	11.89	10.46
Percent of grids (1km) where highways go through	2.97%	4.85%	5.11%
Percent of grids (10km) where highways go through	25.10%	36.82%	39.18%
Population density (person / squared km)	265.88	550.57	393.06
GDP per capita (thousand US dollar)	7.05	13.20	29.00
Average night light	44.74	63.8	74.82
Gini coefficint of GDP	0.53	0.52	0.34
Gini coefficint of GDP per capita	0.29	0.30	0.22

Note: Highways' length density is calculated as the length of highways divided by areas of cities, with unit $km/100km^2$. Highways' area density is calculated as the area of highways (which is the multiplication of length and lane number) divided by areas of cities, with unit $km \times lane/100km^2$. We construct two types of grids covering the cities, one with size $1km \times 1km$, and the other with $10km \times 10km$. Average distance from grids (1km or 10km) to highways is the average value of the distances from each grid's center point to the nearest highways, and percent of grids (1km or 10km) where highways go through is the percentage of grids which intersect with highways. The population, GDP and night light are all gridded dataset: the population comes from LandScan Datasets for 2018, where the resolution is 1 km; the GDP comes from Global Dataset of Gridded Population and GDP (1980-2010 Estimations and 2020-2100 Scenarios) for 2010, where the grid size is 0.5×0.5 degree and the unit is Billion $US\$2005/yr$ under PPP; the night light comes from Flint from China Remote Sensing Satellite Ground Station, Chinese Academy of Science, which is originally from Version 1 VIIRS Day/Night Band Nighttime Lights for 2018 with grid size as 15 arc-second.

The MIDAS data we use for England reports the average traffic speed and flow every 15 minutes at 7,426 measurement sites, which are all on highways. In contrast, G7's GPS data covers the company's trucks on all Chinese roads. We add several restrictions to the two samples to facilitate a direct comparison between China and England. We first restrict our attention to *trucks* or truck-like long vehicles on *highways*. The MIDAS data does not explicitly distinguish trucks from the other vehicles. However, it reports traffic flows conditional on the length of the vehicle. For traffic flows in the MIDAS data, we will only use long vehicles with a length above 6.6 meters, which are arguably the closest counterpart of trucks in G7's GPS records.[45] We then re-construct China's road network to make it entirely based on national and provincial highways.[46] Another important difference is that our GPS records cover four hours a day and two days each month, while MIDAS has data for every 15-minute interval. We extract a sample from the MIDAS data with the same time structure as the GPS data for China. Finally, each observation on speed and traffic flows in the MIDAS data is an average value over 15 minutes.[47] We extract a sample from G7's GPS records that consists of four observations in an hour for each link in China.[48]

We then use the two restricted samples to estimate the same congestion equation. The spatial distribution of the estimated δ_2 is plotted in Figure 15. The distribution of positive and significant δ_2 is plotted in Figure 16. Overall, nearly all links in England (120 out of 122) are congested. The share of congested links in China's highway network is 41%. While the share is higher than 20.97% in our benchmark road network, it is still much lower than the congested share in England.

[45] The MIDAS data groups vehicles into four types by length: (1) less than 5.2 meters, (2) between 5.2 and 6.6 meters, (3) between 6.6 and 11.6 meters, and (4) above 11.6 meters.

[46] This new network is constructed following the same approach as that in Section 2.2, but we replace the NPCC roads by highways. As not all counties are accessible to highways, we skip such counties when constructing the county-pair shortest path.

[47] Recall that the MIDAS data only reports the average speed of all vehicles passing through a measurement site over every 15-minute interval and does not condition the average speed on vehicle type.

[48] We use the 1st, 16th, 31th and 46th minute in each hour.



Figure 15: Comparison of δ_{2kl}s' Spatial Distribution

Note: The red lines indicate significant and positive δ_{2kl}, the orange lines indicate insignificant and positive δ_{2kl}, the yellow lines indicate insignificant and negative δ_{2kl}, and the green lines indicate significant and negative δ_{2kl}. This figure is based on the comparable sample of China and England.

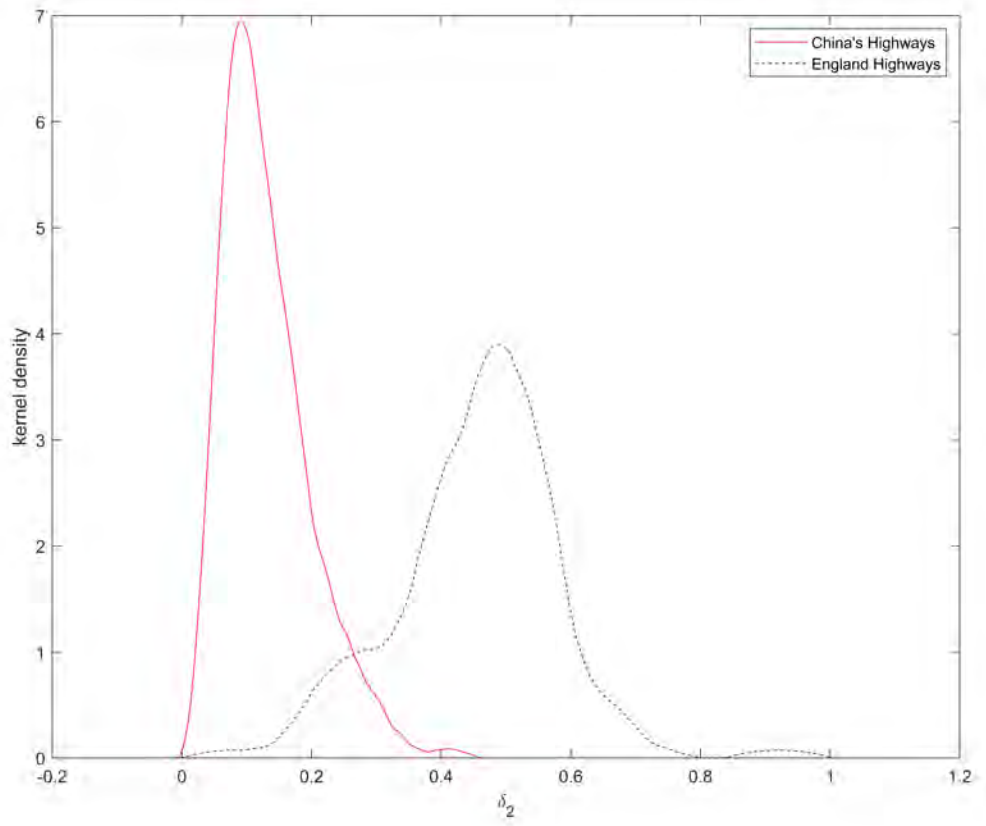


Figure 16: Comparison of δ_{2kl} between China and England

Note: This figure focuses on the positive and significant δ_{2kl}s, and is based on the comparable sample of China and England

A.3 Additional Results on the Fundamental Law of Road Congestion

A.3.1 Link-Level Cross-Sectional Regressions

Table 11: OLS with More Control Variables

Dep: VKT $(\ln)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Capacity $(\ln, lag)$	0.855***	1.033***	0.806***
	(0.060)	(0.117)	(0.071)
Speed limit $(\ln, lag)$	3.215***	0.212	3.720***
	(0.200)	(0.511)	(0.218)
Night light (log)	0.452***	1.848***	0.341***
	(0.105)	(0.251)	(0.122)
Population in 1964 $(\ln)$	0.500***	0.485***	0.531***
	(0.051)	(0.082)	(0.062)
Urban land share in 1980	24.587***	-4.830	28.121***
	(3.529)	(6.401)	(4.222)
Ruggedness	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)
Number of county pairs	-0.001*	-0.001	-0.000
	(0.000)	(0.000)	(0.001)
N	1464	307	1157
R^2	0.635	0.724	0.608
p value for H_0		0.080	

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. H_0: the coefficients of Capacity $(\ln, lag)$ in congested and uncongested groups are the same.

Table 12: 2SLS with Alternative IV, First Stage

Dep: Capacity $(\ln, lag)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Historical road length in 1962 $(\ln)$	0.350***	0.395***	0.324***
	(0.035)	(0.066)	(0.038)
Speed limit $(\ln, lag)$	0.062	-1.098***	0.213
	(0.173)	(0.245)	(0.191)
Night light $(\ln)$	-0.229***	-0.064	-0.259***
	(0.058)	(0.132)	(0.065)
Population in 1964 $(\ln)$	0.228***	0.174***	0.254***
	(0.025)	(0.035)	(0.030)
Urban land share in 1980	7.160***	3.708	4.446*
	(2.480)	(3.914)	(2.680)
Ruggedness index	0.080***	0.018	0.082***
	(0.013)	(0.013)	(0.016)
Number of county pairs	0.004***	0.003***	0.005***
	(0.000)	(0.000)	(0.000)
N	1464	307	1157
R^2	0.689	0.760	0.678
Kleibergen-Paap rk Wald F statistic	114.998	50.550	76.353

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. The IV is the total length of historical road in a buffer of 50 kilometers extended from the city pair (excluding the length within the city pair)

Table 13: 2SLS with Alternative IV, Second Stage

Dep: VMT $(\ln)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Capacity $(\ln, lag)$	0.534***	1.424***	0.376**
	(0.137)	(0.261)	(0.177)
Speed limit $(\ln, lag)$	3.287***	0.589	3.872***
	(0.196)	(0.484)	(0.218)
Night light $(\ln)$	0.322***	1.999***	0.165
	(0.117)	(0.306)	(0.137)
Population in 1964 $(\ln)$	0.599***	0.379***	0.672***
	(0.060)	(0.095)	(0.077)
Urban land share in 1980	26.618***	-6.354	29.812***
	(3.662)	(6.611)	(4.314)
Ruggedness	0.140***	0.067***	0.177***
	(0.019)	(0.026)	(0.024)
Number of county pairs	0.001	-0.002**	0.002**
	(0.001)	(0.001)	(0.001)
N	1464	307	1157
R^2	0.624	0.709	0.589
p value for H_0		0.005	

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. The IV is the total length of historical road in a buffer of 50 kilometers extended from the city pair (excluding the length within the city pair) H_0: the coefficients of Capacity $(\ln, lag)$ in congested and uncongested groups are the same.

Table 14: Indirect Inference with Alternative IV, Second Stage

Dep: VMT $(\ln)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Capacity $(\ln, lag)$	0.793***	1.353***	0.753***
	(0.119)	(0.221)	(0.152)
Speed limit $(\ln, lag)$	1.932***	0.640	2.189***
	(0.166)	(0.518)	(0.181)
Night light $(\ln)$	0.233**	1.604***	0.144
	(0.098)	(0.299)	(0.117)
Population in 1964 $(\ln)$	0.728***	0.506***	0.766***
	(0.054)	(0.086)	(0.068)
Urban land share in 1980	28.431***	6.115	30.395***
	(3.079)	(7.077)	(3.683)
Ruggedness	0.016	0.030	0.022
	(0.017)	(0.025)	(0.021)
Number of county pairs	-0.000	-0.001*	0.001
	(0.001)	(0.001)	(0.001)
N	1464	307	1157
R^2	0.729	0.775	0.709
p value for H_0		0.005	

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. The IV is the total length of historical road in a buffer of 50 kilometers extended from the city pair (excluding the length within the city pair). H_0: the coefficients of Capacity $(\ln, lag)$ in congested and uncongested groups are the same.

A.3.2 Road-Level Cross-Sectional Regressions

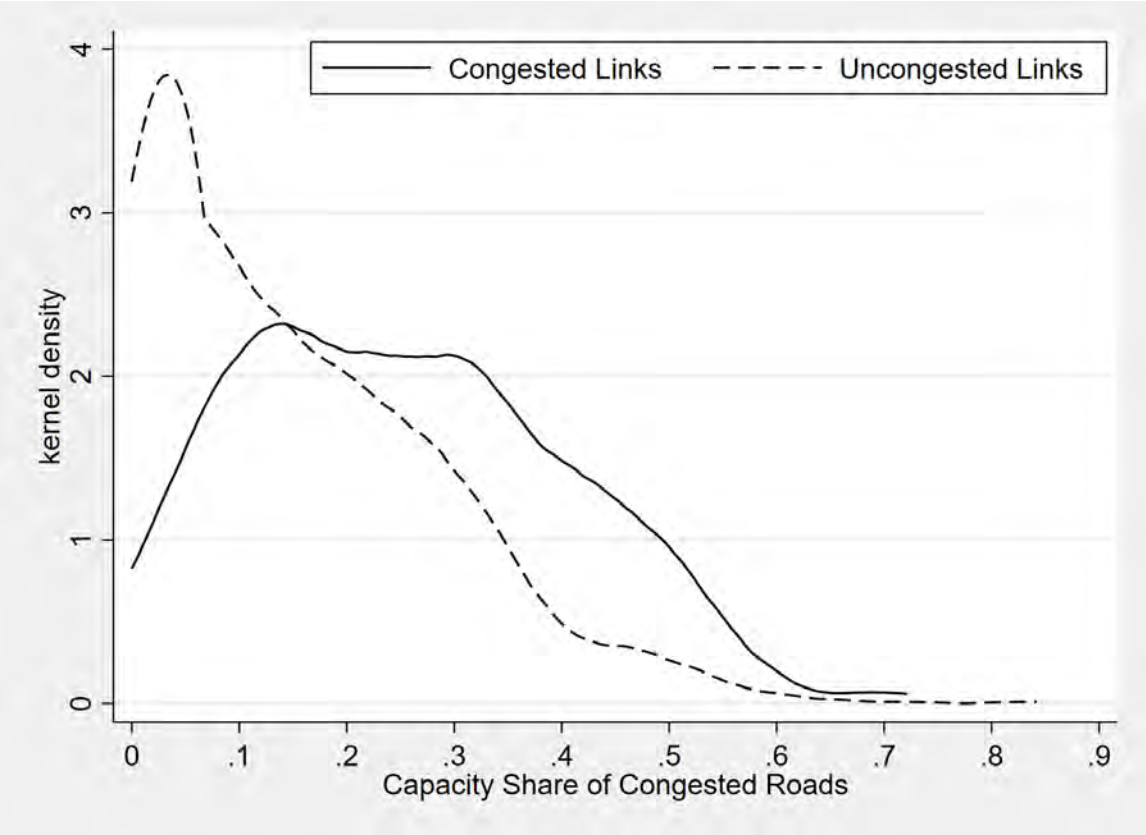


Figure 17: Heterogeneity within Link

Table 15: The Longest Provincial or National Road

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	OLS, Dep: VKT (ln)			First-Stage, Dep: Capacity (ln, lag)			2SLS, Dep: VKT (ln)		
	All Links	Congested Links	Uncongested Links	All Links	Congested Links	Uncongested Links	All Links	Congested Links	Uncongested Links
Capacity (ln, lag)	0.631***	0.821***	0.588***				0.552***	0.801**	0.519***
	(0.034)	(0.083)	(0.035)				(0.082)	(0.319)	(0.081)
Historical road length in 1962 (ln)				0.187***	0.150***	0.193***			
				(0.016)	(0.040)	(0.017)			
Speed limit (ln, lag)	2.582***	3.108***	2.480***	0.796***	0.880***	0.787***	2.649***	3.125***	2.537***
	(0.068)	(0.205)	(0.073)	(0.049)	(0.127)	(0.054)	(0.097)	(0.364)	(0.098)
Night light (ln)	0.362***	0.429**	0.308***	-0.203***	-0.206	-0.215***	0.359***	0.427**	0.306***
	(0.085)	(0.208)	(0.092)	(0.066)	(0.145)	(0.075)	(0.084)	(0.207)	(0.091)
Population in 1964 (ln)	0.258***	0.323***	0.226***	0.042	0.080	0.017	0.260***	0.324***	0.226***
	(0.042)	(0.098)	(0.049)	(0.041)	(0.100)	(0.047)	(0.042)	(0.094)	(0.049)
Urban land share in 1980	13.757***	21.203**	9.921**	1.927	9.385	-1.685	13.269***	21.221**	9.254**
	(3.914)	(10.691)	(4.329)	(3.042)	(7.967)	(3.304)	(4.009)	(10.466)	(4.457)
Ruggedness index	0.072***	0.151***	0.056***	0.092***	0.090**	0.087***	0.081***	0.153***	0.063***
	(0.017)	(0.053)	(0.018)	(0.014)	(0.035)	(0.015)	(0.019)	(0.053)	(0.020)
Number of county pairs	0.000	-0.000	0.000	0.001***	0.001***	0.001***	0.000	-0.000	0.000
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
N	1450	258	1192	1450	258	1192	1450	258	1192
R^2	0.685	0.679	0.679	0.350	0.302	0.358	0.684	0.679	0.678
F $statistic$				135.965	14.031	125.562			
p value for H_0		0.002						0.086	

Note: Robust standard errors in parentheses * p < 0.10, ** p < 0.05, *** p < 0.01. The sample is the cross-section of 2018. Here "congestion" is estimated based on data of 2018. This is constructed by following the similar logic to that in link-level analysis: we first construct 10km and 1km buffering areas along each road and calculate the total length of historical roads within them respectively, and then the length within the 10km buffering area subtracting the length within the 1km buffering area becomes our instrument. The length within the 1km buffering area is subtracted to avoid serial correlation. H_0: the coefficients of Capacity (ln, lag) in congested and uncongested groups are the same.

A.3.3 Panel Regressions

We next run panel fixed effects regressions to control for unobserved and time-invariant link characteristics. The panel data has a few features that deserve discussions. First, due to the rapid expansion of China's road network (Qin, 2016; He et al., 2020), the average number of lanes across city-pair links increased by 5.87% in our sample period between 2013 and 2018. This leads to substantial over-time variations in link capacity that can be used to test the fundamental law of road congestion. Second, truck flows by GPS records increased dramatically. A significant proportion of the increase came from the expansion of G7's market share. Therefore, over-time variations in truck flows in our sample is a contaminated measure of the actual variations. To overcome the difficulty, we look at truck speed, which turns out to be stable over time and should be immune to the company's business expansion (see Figure 12 and 13). In fact, the relationship between link capacity and driving speed is the other side of the fundamental law of road congestion, which predicts that driving speed increases immediately after the expansion for congested links. Moreover, the effect will decay over time and disappear eventually (e.g., see Hsu and Zhang, 2014). The dynamic feature of the data allows us to keep track of the effect of expanding link capacity over time.

We first run the following regression:

$$\ln\left(speed_{klti}\right) = \gamma_{1i} + \gamma_{2i} \times \ln\left(cap_{klt-1}\right) + \rho X_{kl\iota} + FE_{kl} + FE_t + \epsilon_{klti}. \tag{34}$$

Since we can access China's navigation road map at the end of year 2013, 2015 and 2017, cap_{klt-1} denotes the capacity of link (k,l) at the beginning of year t, where $t \in \{2014, 2016, 2018\}$. $i \in [3, 12]$ represents the month in the following year. We drop February to exclude the influence of the Spring Festival. So, $speed_{klti}$ refers to the average driving speed on link (k,l) in the $i-$th month of year t. $X_{kl\iota}$ represents a vector of control variables, some of which are measured at $\iota = t-1$ like speed limit, and others are measured at $\iota = t-2$ like coefficient of variation and mean value of night light as will be discussed below. We run the panel regression for each month (except February) and plot the estimated γ_{2i} in Figure 18. We find a significant increase in driving speed for both congested and uncongested links. The increase is larger and more persistent for congested links. Consistent with the fundamental law of road congestion, the increase becomes statistically insignificant in the second half of the year. The estimates are even negative for uncongested links.

One concern when estimating the effect of road capacity on traffic is that road capacity is endogenous, for example if new lanes were built where high traffic suggests that there is a strong demand for new road capacity. In the panel fixed effects setting, time-invariant factors are absorbed by the fixed effects, but there may be unobserved shocks that affect demand and supply. The typical instruments used in the literature for transportation infrastructure investments, for example historical road networks, are time-invariant and therefore cannot

directly be used in this panel analysis.[49] We propose a novel time-varying instrumental variable that is purely driven by road supply factors.

The idea of the instrument is that along a link between two cities, certain segments become supply-driven bottlenecks. These bottlenecks are supply-driven in the sense that the number of lanes of the individual segment is below the other segments on the same link. Subsequent infrastructure investments likely attempt to alleviate these bottlenecks. Consequently, links with currently more heterogeneous road widths across the segments on the same link are more likely to have an expansion of the capacity in the following year.

We measure the heterogeneity in the width along a link, and thus the likelihood of future capacity upgrades along that link, as the coefficient of variation of the number of roads, denoted by CV_{klt}. More precisely, we first compute the CV of the number of lanes across segments for each road on a link and then obtain the average of CV across roads for the link, weighted by road area.[50] We argue that the IV measures the severity of bottlenecks in a link and, hence, the lagged CV_{klt-2} predicts the increase in $lane_{klt-1}$. We include the lagged capacity in the instruments, such that the lagged CV and lagged capacity together predict the current link capacity.

The exclusion restriction is that the variation in the number of lanes affects changes in vehicle-km-traveled only through changes in capacity. The instrument should therefore only capture supply-side determinants of traffic. One potential threat to identification is that a link's variation in number of lanes also reflects the variation in economic conditions of the area that the road passes through, and thus potential drivers of demand. For example, a link that connects certain high-density areas (with more lanes) but passes through low-density areas (with fewer lanes) along the way will have a high coefficient of variation, which thus also reflects the economic inequality. The time-invariant part of these factors is absorbed by the link fixed effects.[51] But if there are shocks to the dispersion in economic activity that also affect the demand side of travel, then the exclusion restriction would not hold. We therefore control for the variation in economic density over time along the path of the road.

We run the 2SLS regression through the following specification

$$\ln\left(cap_{klt-1}\right) = d_{1i} + d_{2i} \times \ln\left(CV_{klt-2}\right) + d_{3i} \times \ln\left(cap_{klt-2}\right) + \rho X_{kl\iota} + FE_{kl} + FE_t + \epsilon_{klti}, \tag{35}$$

$$\ln\left(speed_{klti}\right) = \gamma_{1i} + \gamma_{2i} \times \ln\left(\widehat{cap_{klt-1}}\right) + \varrho X_{kl\iota} + FE_{kl} + FE_t + \varepsilon_{klti}, \tag{36}$$

[49] See the discussion in Section 1.1.

[50] Each link consists of several roads, where a road is defined by the official name. Each road is a group of segments, which are typically separated from each other by intersection. The typical road is 100-200 km long and consists of 10-20 segments.

[51] The link fixed effects absorb the initial economic conditions in each node and in each link.

and find a strong first stage for both congested and uncongested links. The results are reported in Table 16. The second-stage results are plotted in Figure 19. The IV approach reveals larger differences between congested and uncongested links. Almost all of the monthly estimates of γ_{2i} are insignificant for uncongested links. In contrast, the estimates exhibit a clear downward trend for congested links. The estimates in January, March and April are not only significant, but at least three times as large as their counterparts for uncongested links. The estimates move down in both significance and magnitude in May and afterwards, implying that more traffic be attracted to the expanded links. This is exactly what the fundamental law would predict.

Table 16: Panel Results, First Stage

Dep: Capacity $(\ln)$	(1) All Links	(2) Congested Links	(3) Uncongested Links
Link CV $(\ln, lag)$	0.028***	0.031**	0.028***
	(0.006)	(0.012)	(0.007)
Capacity $(\ln, lag)$	0.205***	0.201***	0.207***
	(0.019)	(0.047)	(0.021)
Speed limit $(\ln, lag)$	0.361***	0.339***	0.366***
	(0.014)	(0.036)	(0.016)
CV of night light along the link $(\ln, lag)$	-0.012	0.006	-0.016
	(0.011)	(0.015)	(0.012)
Mean of night light within the city pair $(\ln, lag)$	0.152	0.242*	0.111
	(0.106)	(0.132)	(0.135)
Link FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	4382	921	3461
$Within\ R^2$	0.409	0.455	0.395
$F\ statistic$	228.13	51.95	179.67

Note: Robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. The sample covers year 2013, 2015, 2017. Here "congestion" is estimated based on data of 2018.

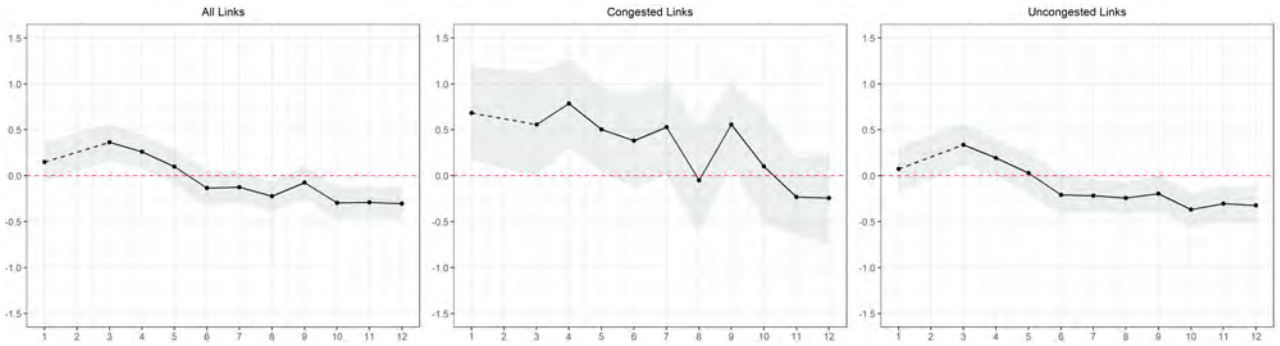


Figure 18: Panel FE, by month

Note: The vertical axis indicates regression coefficients of Capacity ($\ln, lag$), and the horizontal axis indicates month. The shade areas are 95% confidence intervals.

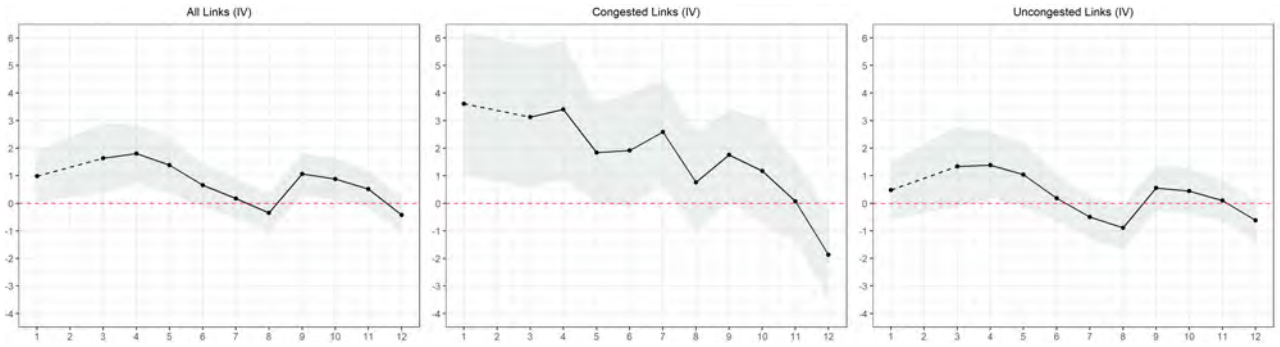


Figure 19: Panel 2SLS, by month

Note: The vertical axis indicates regression coefficients of Capacity ($\ln, lag$), and the horizontal axis indicates month. The shade areas are 95% confidence intervals.

A.4 Algorithm of the Structural Estimation

Our estimation of θ, κ and ζ follows this algorithm.

- Step 1: Start with a guess for $hours_{kl}$, denoted by $hours_{kl}^{guess}$. Then, we obtain t_{kl} and τ_{ij} from (26) and (22), respectively.
- Step 2: We use city-level GDP, denoted by Y_i^{data}, for Y_i. The market clearing and balanced trade conditions, (12) and (13), solve p_i and P_i. Then, (10) gives $X_{ij} = \tau_{ij}^{1-\sigma} p_i^{1-\sigma} P_j^{\sigma-1} Y_j$. Substituting X_{ij} into (24), we obtain Ξ_{kl}.
- Step 3: Substituting Ξ_{kl} into (25), we obtain $flow_{kl}$, which generate $traf_{kl}$ through (1) and, then, $hours_{kl}^{updated}$ by (2).
- Step 5: Iterate on $hours_{kl}^{updated}$ until they convergence. $traf_{kl}^{sim}$ and $hours_{kl}^{sim}$ are $traf_{kl}$ and $hours_{kl}$ associated with the converged $hours_{kl}$.

A.5 Welfare Implications, Additional Figures and Tables

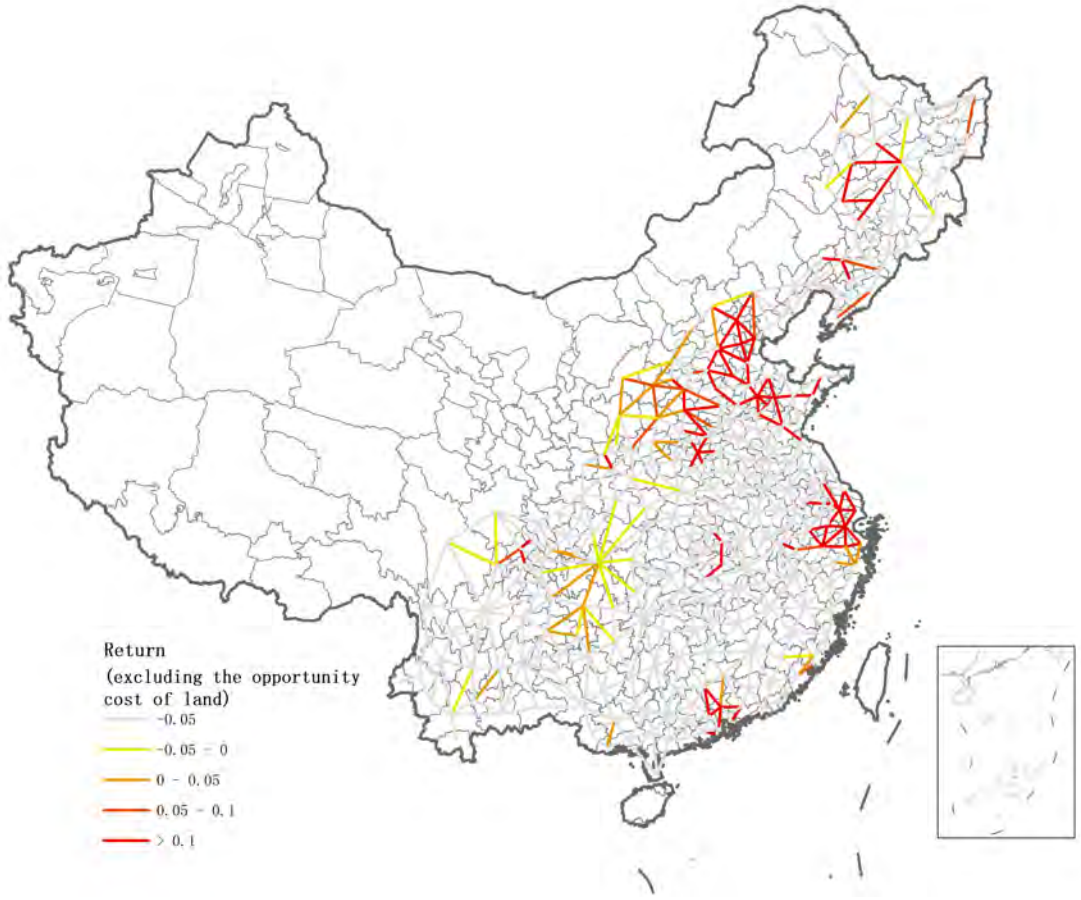


Figure 20: Spatial Distribution of Returns when Excluding the Opportunity Cost of Land

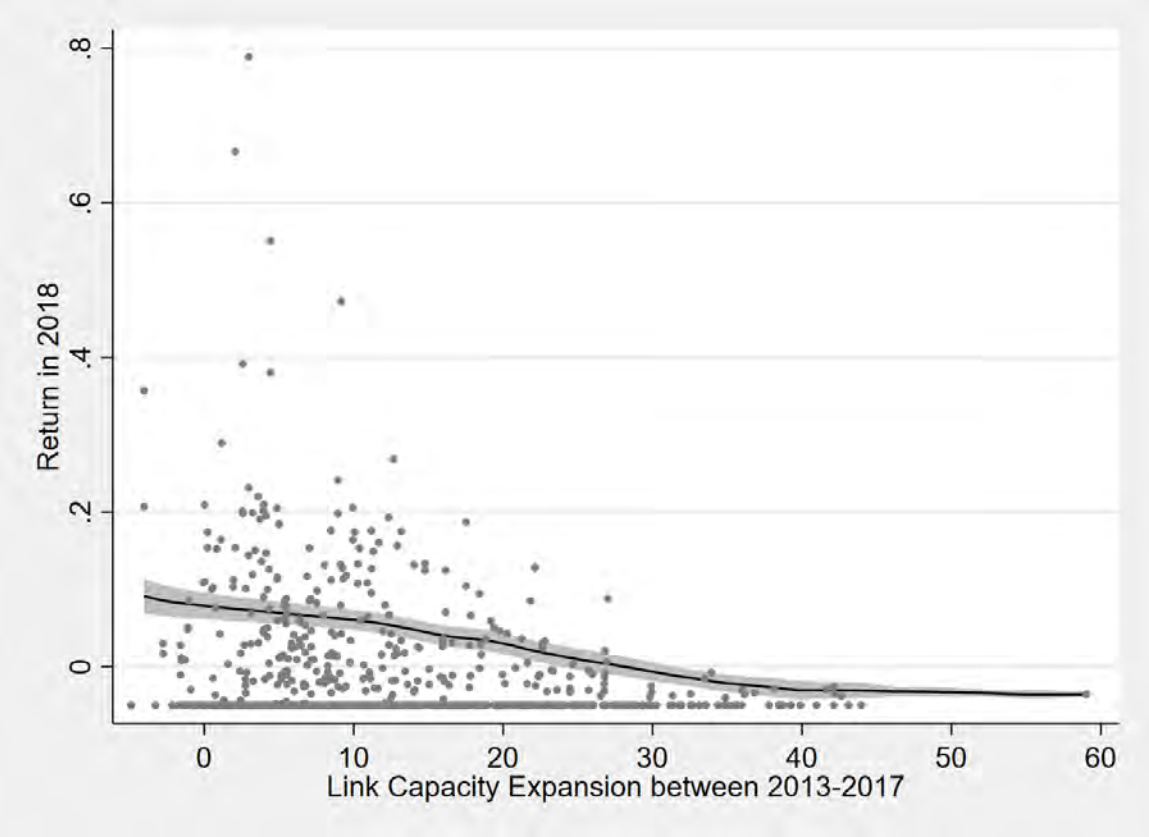


Figure 21: Returns and Link Capacity Expansion

Note: The black line indicates the polynomial fitted line for congested links, and the shaded area represents the corresponding 95% confidence interval.

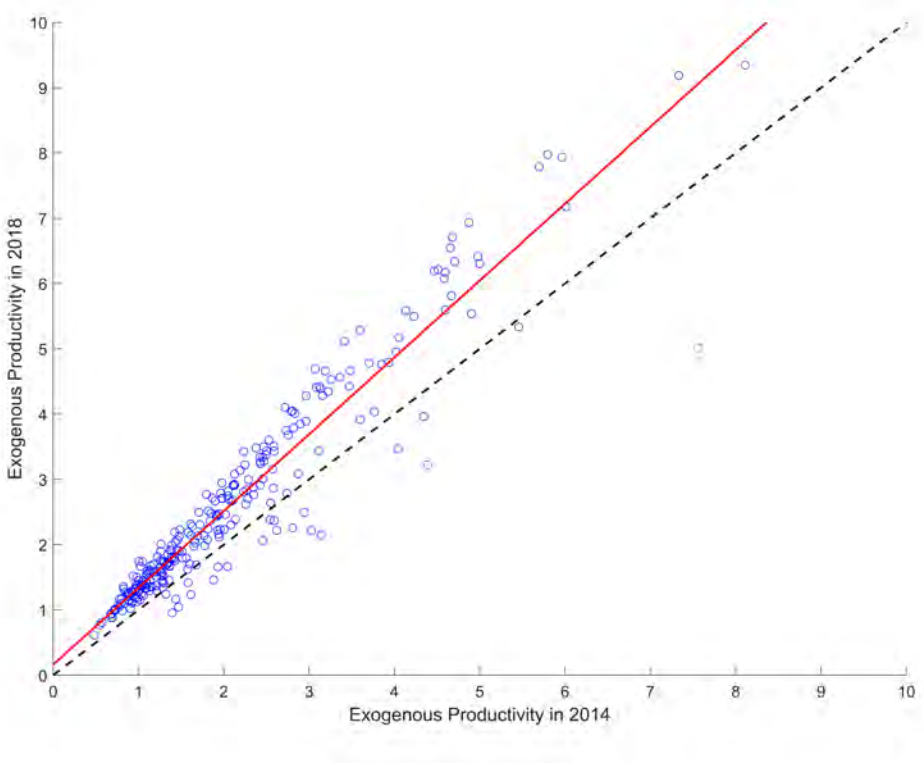


Figure 22: Exogenous Productivity

Note: The black-dashed line is 45 degree line and the red-solid line is the polynomial fitted line.

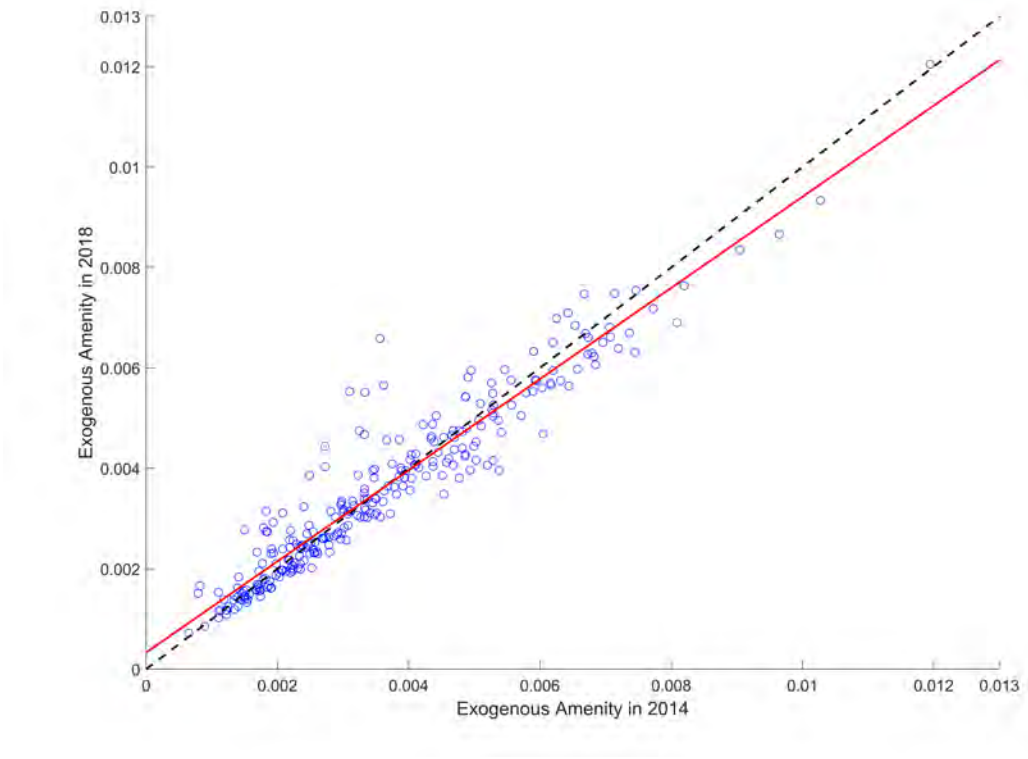


Figure 23: Exogenous Amenity

Note: The black-dashed line is 45 degree line and the red-solid line is the polynomial fitted line.

Table 17: Top 10 Links in Terms of Return, China

(1)	(2)	(3)	(4)	(5)
Rank	Including Land Cost		Excluding Land Cost	
	Link	Return	Link	Return
1	Wuxi-Suzhou	0.789	Shanghai-Suzhou	9.814
2	Shanghai-Jiaxing	0.666	Shanghai-Jiaxing	9.385
3	Shanghai-Suzhou	0.551	Suzhou-Shanghai	7.013
4	Xinxiang-Jiaozuo	0.472	Wuxi-Suzhou	5.673
5	Foshan-Guangzhou	0.392	Foshan-Guangzhou	5.291
6	Suzhou-Shanghai	0.381	Shanghai-Nantong	4.163
7	Xuchang-Zhengzhou	0.357	Tianjin-Langfang	3.727
8	Binzhou-Zibo	0.289	Guangzhou-Foshan	2.951
9	Weifang-Zibo	0.268	Nantong-Shanghai	2.929
10	Xinxiang-Zhengzhou	0.241	Jiaxing-Shanghai	2.635

Note: In this table we focus on the case when $\lambda = 1$.

Table 18: Top 10 Links in terms of Gain, England

Rank		Link		Benefit / GDP (‰)
(1)	Derbyshire and Nottinghamshire	→	Bedfordshire and Hertfordshire	0.4791
(2)	Leicestershire, Rutland and Northamptonshire	→	Bedfordshire and Hertfordshire	0.4643
(3)	Leicestershire, Rutland and Northamptonshire	→	Derbyshire and Nottinghamshire	0.4546
(4)	Bedfordshire and Hertfordshire	→	Leicestershire, Rutland and Northamptonshire	0.4449
(5)	Herefordshire, Worcestershire and Warwickshire	→	West Midlands	0.3595
(6)	West Midlands	→	Herefordshire, Worcestershire and Warwickshire	0.3474
(7)	Hampshire and Isle of Wight	→	Surrey, East and West Sussex	0.2694
(8)	Bedfordshire and Hertfordshire	→	Outer London - West and North West	0.2664
(9)	Tees Valley and Durham	→	Northumberland and Tyne and Wear	0.2576
(10)	West Yorkshire	→	Greater Manchester	0.2552

Note: In this table we focus on the case when $\lambda = 1$.

A.6 Derivations of Welfare Elasticities

A.6.1 Elasticity of Welfare to Expected Trade Cost

In this section we rephrase the whole process from Allen and Arkolakis (2019) to derive the formula of $d\ln W/d\ln \tau kl$

According to Allen and Arkolakis (2019), we first express the equilibrium conditions by explicitly incorporating welfare equalization as

$$w_i^{\sigma} L_i^{1-\alpha(\sigma-1)} = W^{1-\sigma} \sum_j \tau_{ij}^{1-\sigma} \bar{A}_i^{\sigma-1} \bar{u}_j^{\sigma-1} w_j^{\sigma} L_j^{1+\beta(\sigma-1)}$$
$$w_i^{1-\sigma} L_i^{\beta(1-\sigma)} = W^{1-\sigma} \sum_j \tau_{ji}^{1-\sigma} \bar{A}_j^{\sigma-1} \bar{u}_i^{\sigma-1} w_j^{1-\sigma} L_j^{\alpha(\sigma-1)}$$

and define

$$x_i \equiv \left(W^{\frac{-1}{\alpha+\beta}} L_i\right)^{\beta(\sigma-1)+1} w_i^{\sigma}$$
$$y_i \equiv \left(W^{\frac{-1}{\alpha+\beta}} L_i\right)^{\alpha(\sigma-1)} w_i^{1-\sigma}$$
$$K_{ij} \equiv \tau_{ij}^{1-\sigma} \bar{A}_i^{\sigma-1} \bar{u}_j^{\sigma-1}$$

which tell the equilibrium conditions can be expressed as

$$x_i^{\frac{\alpha+1}{\beta(\sigma-1)+1+\alpha\sigma}} y_i^{\frac{-\beta+1}{\beta(\sigma-1)+1+\alpha\sigma}} = \sum_j K_{ij} x_j y_i$$
$$x_i^{\frac{\alpha+1}{\beta(\sigma-1)+1+\alpha\sigma}} y_i^{\frac{-\beta+1}{\beta(\sigma-1)+1+\alpha\sigma}} = \sum_j K_{ji} x_i y_j$$

Then by applying the implicit function theorem on equilibrium conditions and normalizing $w_s = 1$,[52] we can get

$$\begin{aligned}
\frac{d\ln y_i}{d\ln \tau_{kl}} &= (\sigma-1)\left[DGI_{i,k}\frac{X_{kl}}{Y_k} + DGI_{i,N+l}\frac{X_{kl}}{E_l}\right] \\
&= (\sigma-1)\left[DGI_{i,k}\frac{X_{kl}}{Y_k}\right] \ if\ l = s \\
\frac{d\ln x_i}{d\ln \tau_{kl}} &= (\sigma-1)\left[DGI_{N+i,k}\frac{X_{kl}}{Y_k} + DGI_{N+i,N+l}\frac{X_{kl}}{E_l}\right] \\
&= (\sigma-1)\left[DGI_{N+i,k}\frac{X_{kl}}{Y_k}\right] \ if\ l = s
\end{aligned}$$

[52] s indicates the area with the highest GDP.

where

$$DG \equiv \begin{bmatrix} \frac{\sigma(\alpha+\beta)}{\beta(\sigma-1)+1+\alpha\sigma} I & \frac{X}{Y} - \frac{1+\alpha}{\beta(\sigma-1)+1+\alpha\sigma} I \\ \left(\frac{X}{E}\right)' - \frac{1-\beta}{\beta(\sigma-1)+1+\alpha\sigma} I & \frac{(\sigma-1)(\alpha+\beta)}{\beta(\sigma-1)+1+\alpha\sigma} I \end{bmatrix} - \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{X_{1s}}{E_s} & \cdots & \frac{X_{ss}}{E_s} + \frac{\beta}{\alpha} & \cdots & \frac{X_{Ns}}{E_s} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$DGI \equiv DG^{-1}$$

$$I \equiv \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\frac{X}{Y} \equiv \begin{bmatrix} \frac{X_{11}}{Y_1} & \frac{X_{12}}{Y_1} & \cdots & \frac{X_{1N}}{Y_1} \\ \frac{X_{21}}{Y_2} & \frac{X_{22}}{Y_2} & \cdots & \frac{X_{2N}}{Y_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{X_{N1}}{Y_N} & \frac{X_{N2}}{Y_N} & \cdots & \frac{X_{NN}}{Y_N} \end{bmatrix}, \frac{X}{E} \equiv \begin{bmatrix} \frac{X_{11}}{E_1} & \frac{X_{12}}{E_1} & \cdots & \frac{X_{1N}}{E_1} \\ \frac{X_{21}}{E_2} & \frac{X_{22}}{E_2} & \cdots & \frac{X_{2N}}{E_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{X_{N1}}{E_N} & \frac{X_{N2}}{E_N} & \cdots & \frac{X_{NN}}{E_N} \end{bmatrix}$$

Finally, based on the definitions of x_i and y_i, we can get

$$\frac{d\ln W}{d\ln \tau_{kl}} = -\frac{2-\rho}{2\sigma-1}\left(\frac{dx_i}{d\tau_{kl}} + \frac{\sigma}{\sigma-1}\frac{dy_i}{d\tau_{kl}}\right) + (\alpha+\beta)\frac{d\ln L_i}{d\ln \tau_{kl}}$$

Further utilizing that $\sum_i L_i = \bar{L}$,we can obtain [53]

$$\begin{aligned} \frac{d\ln W}{d\ln \tau_{kl}} &= -\frac{2-\rho}{2\sigma-1}\sum_i \frac{L_i}{\bar{L}}\left(\frac{dx_i}{d\tau_{kl}} + \frac{\sigma}{\sigma-1}\frac{dy_i}{d\tau_{kl}}\right) \\ &= \begin{cases} -X_{kl}\left(1+\varsigma_k+\varpi_l\right) \\ -X_{kl}\left(1+\varsigma_k\right) \qquad if\ l=s \end{cases} \end{aligned}$$

[53]

$$\begin{aligned} &\frac{d\ln W}{d\ln \tau_{kl}} = -\frac{2-\rho}{2\sigma-1}\left(\frac{dx_i}{d\tau_{kl}} + \frac{\sigma}{\sigma-1}\frac{dy_i}{d\tau_{kl}}\right) + (\alpha+\beta)\frac{d\ln L_i}{d\ln \tau_{kl}} \\ \Longleftrightarrow &\frac{L_i}{\bar{L}}\frac{d\ln W}{d\ln \tau_{kl}} == -\frac{2-\rho}{2\sigma-1}\frac{L_i}{\bar{L}}\left(\frac{dx_i}{d\tau_{kl}} + \frac{\sigma}{\sigma-1}\frac{dy_i}{d\tau_{kl}}\right) + (\alpha+\beta)\frac{1}{\bar{L}}\frac{dL_i}{d\ln \tau_{kl}} \\ \Longrightarrow &\frac{d\ln W}{d\ln \tau_{kl}}\sum_i \frac{L_i}{\bar{L}} = -\frac{2-\rho}{2\sigma-1}\sum_i \frac{L_i}{\bar{L}}\left(\frac{dx_i}{d\tau_{kl}} + \frac{\sigma}{\sigma-1}\frac{dy_i}{d\tau_{kl}}\right) + (\alpha+\beta)\frac{1}{\bar{L}}\frac{\sum_i dL_i}{d\ln \tau_{kl}} \\ \Longrightarrow &\frac{d\ln W}{d\ln \tau_{kl}} = -\frac{2-\rho}{2\sigma-1}\sum_i \frac{L_i}{\bar{L}}\left(\frac{dx_i}{d\tau_{kl}} + \frac{\sigma}{\sigma-1}\frac{dy_i}{d\tau_{kl}}\right) \end{aligned}$$

where

$$\rho \equiv \frac{2+\alpha-\beta}{\beta(\sigma-1)+1+\alpha\sigma}$$
$$\varsigma_k \equiv \frac{(\alpha+\beta)(\sigma-1)}{\beta(\sigma-1)+1+\alpha\sigma}\sum_i \frac{L_i}{\bar{L}}\left(DGI_{N+i,k}+\frac{\sigma}{\sigma-1}DGI_{i,k}\right)\frac{1}{Y_k}-1$$
$$\varpi_l \equiv \frac{(\alpha+\beta)(\sigma-1)}{\beta(\sigma-1)+1+\alpha\sigma}\sum_i \frac{L_i}{\bar{L}}\left(DGI_{N+i,N+l}+\frac{\sigma}{\sigma-1}DGI_{i,N+l}\right)\frac{1}{E_l}$$

Meanwhile, still based on the definitions of x_i and y_i, we can get

$$\frac{d\ln w_i}{d\ln \tau_{kl}} = \frac{\alpha}{\beta(\sigma-1)+1+\alpha\sigma}\frac{d\ln x_i}{d\ln \tau_{kl}} - \frac{1}{\sigma-1}\frac{\beta(\sigma-1)+1}{\beta(\sigma-1)+1+\alpha\sigma}\frac{d\ln y_i}{d\ln \tau_{kl}}$$
$$\frac{d\ln L_i}{d\ln \tau_{kl}} = \frac{1}{\beta(\sigma-1)+1}\frac{d\ln x_i}{d\ln \tau_{kl}} - \frac{\sigma}{\beta(\sigma-1)+1}\frac{d\ln w_i}{d\ln \tau_{kl}} + \frac{1}{\alpha+\beta}\frac{d\ln W}{d\ln \tau_{kl}}$$
$$\frac{d\ln p_i}{d\ln \tau_{kl}} = \frac{d\ln w_i}{d\ln \tau_{kl}} - \alpha\frac{d\ln L_i}{d\ln \tau_{kl}}$$
$$\frac{d\ln P_i}{d\ln \tau_{kl}} = \frac{d\ln w_i}{d\ln \tau_{kl}} - \frac{d\ln W}{d\ln \tau_{kl}} + \beta\frac{d\ln L_i}{d\ln \tau_{kl}}$$
$$\frac{d\ln Y_i}{d\ln \tau_{kl}} = \frac{d\ln w_i}{d\ln \tau_{kl}} + \frac{d\ln L_i}{d\ln \tau_{kl}}$$
$$\frac{d\ln X_{ij}}{d\ln \tau_{kl}} = (1-\sigma)\,1\{ij=kl\} + (1-\sigma)\frac{d\ln p_i}{d\ln \tau_{kl}} + (\sigma-1)\frac{d\ln P_j}{d\ln \tau_{kl}} + \frac{d\ln Y_j}{d\ln \tau_{kl}}$$

A.6.2 Elasticity of Welfare to Expected Trade Cost without Externality

When there is no externality, the market equilibrium is equivalent to social planner's solution, and thus we can directly solve for $d\ln W_i/d\ln \tau_{kl}$ through social welfare maximization and Envelop Theorem

$$\frac{d\ln W}{d\ln \tau_{kl}} = -\frac{X_{kl}}{\sum_i\sum_j X_{ij}}$$

Some related elasticities can also be derived as follows. The simplified equilibrium conditions become

$$w_i^\sigma L_i = W^{1-\sigma}\sum_j \tau_{ij}^{1-\sigma}\bar{A}_i^{\sigma-1}\bar{u}_j^{\sigma-1}w_j^\sigma L_j$$
$$w_i^{1-\sigma} = W^{1-\sigma}\sum_j \tau_{ji}^{1-\sigma}\bar{A}_j^{\sigma-1}\bar{u}_i^{\sigma-1}w_j^{1-\sigma}.$$

Let's define

$$x_i \equiv w_i^\sigma L_i$$
$$y_i \equiv w_i^{1-\sigma}$$
$$K_{ij} \equiv \tau_{ij}^{1-\sigma}\bar{A}_i^{\sigma-1}\bar{u}_j^{\sigma-1}$$

and then transform the equilibrium conditions into

$$x_i = W^{1-\sigma} \sum_j K_{ij} x_j$$
$$y_i = W^{1-\sigma} \sum_j K_{ji} y_j$$

Then based on the formula of $d \ln W / d \ln \tau_{kl}$ and by applying the implicit function theorem on equilibrium conditions, we can get

$$\left[\frac{X}{Y} - I\right] \begin{bmatrix} \frac{d \ln x_1}{d \ln \tau_{kl}} \\ \frac{d \ln x_2}{d \ln \tau_{kl}} \\ \vdots \\ \frac{d \ln x_N}{d \ln \tau_{kl}} \end{bmatrix} + (\sigma - 1) \frac{X_{kl}}{\sum_i \sum_j X_{ij}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + (1 - \sigma) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{X_{kl}}{Y_k} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

$$\left[\left(\frac{X}{E}\right)' - I\right] \begin{bmatrix} \frac{d \ln y_1}{d \ln \tau_{kl}} \\ \frac{d \ln y_2}{d \ln \tau_{kl}} \\ \vdots \\ \frac{d \ln y_N}{d \ln \tau_{kl}} \end{bmatrix} + (\sigma - 1) \frac{X_{kl}}{\sum_i \sum_j X_{ij}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + (1 - \sigma) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{X_{kl}}{E_l} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

where X_{kl}/Y_k is located in the $k-th$ row, and X_{kl}/E_l is located in the $l-th$ row. Because $\left[\left(\frac{X}{E}\right)' - I\right]$ is singular, we need to add normalization (e.g. $w_s = 1 \Longleftrightarrow y_s = 1$) and then obtain

$$\frac{d \ln y_s}{d \ln \tau_{kl}} = 0$$

$$\begin{bmatrix} \frac{d \ln y_1}{d \ln \tau_{kl}} \\ \vdots \\ \frac{d \ln y_{s-1}}{d \ln \tau_{kl}} \\ \frac{d \ln y_{s+1}}{d \ln \tau_{kl}} \\ \vdots \\ \frac{d \ln y_N}{d \ln \tau_{kl}} \end{bmatrix} = \left[\left(\frac{X}{E}\right)' - I\right]^{-1}_{-s,-s} \left\{ - (\sigma - 1) \frac{X_{kl}}{\sum_i \sum_j X_{ij}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{-s} - (1 - \sigma) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{X_{kl}}{E_l} \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{-s} \right\}$$

$$\Rightarrow \frac{d \ln w_i}{d \ln \tau_{kl}} = \frac{1}{1 - \sigma} \frac{d \ln y_i}{d \ln \tau_{kl}}$$

where $-s$ indicates drop the $s-th$ row or column. Because $\left[\frac{X}{Y}-I\right]$ is also singular, we first normalize $L_s=1$ to get

$$\frac{d\ln L_s}{d\ln \tau_{kl}}=0$$

$$\begin{bmatrix}\frac{d\ln L_1}{d\ln \tau_{kl}}\\ \vdots \\ \frac{d\ln L_{s-1}}{d\ln \tau_{kl}}\\ \frac{d\ln L_{s+1}}{d\ln \tau_{kl}}\\ \vdots \\ \frac{d\ln L_N}{d\ln \tau_{kl}}\end{bmatrix}=\left[\frac{X}{Y}-I\right]^{-1}_{-s,-s}\left\{-(\sigma-1)\frac{X_{kl}}{\sum_i\sum_j X_{ij}}\begin{bmatrix}1\\1\\ \vdots \\1\end{bmatrix}_{-s}-(1-\sigma)\begin{bmatrix}0\\ \vdots \\0\\ \frac{X_{kl}}{Y_k}\\0\\ \vdots \\0\end{bmatrix}_{-s}-\sigma\left[\frac{X}{Y}-I\right]\begin{bmatrix}\frac{d\ln w_1}{d\ln \tau_{kl}}\\ \vdots \\ \frac{d\ln w_{s-1}}{d\ln \tau_{kl}}\\ \frac{d\ln w_{s+1}}{d\ln \tau_{kl}}\\ \vdots \\ \frac{d\ln w_N}{d\ln \tau_{kl}}\end{bmatrix}\right\}$$

and then recover the scale of labor elasticity using $\sum_i L_i=\bar{L}$. Thus we can get

$$\frac{d\ln x_i}{d\ln \tau_{kl}}=\sigma\frac{d\ln w_i}{d\ln \tau_{kl}}+\frac{d\ln L_i}{d\ln \tau_{kl}}$$

Meanwhile, we can get

$$\begin{aligned}\frac{d\ln p_i}{d\ln \tau_{kl}}&=\frac{d\ln w_i}{d\ln \tau_{kl}}\\ \frac{d\ln P_i}{d\ln \tau_{kl}}&=\frac{d\ln w_i}{d\ln \tau_{kl}}-\frac{d\ln W}{d\ln \tau_{kl}}\\ \frac{d\ln Y_i}{d\ln \tau_{kl}}&=\frac{d\ln w_i}{d\ln \tau_{kl}}+\frac{d\ln L_i}{d\ln \tau_{kl}}\\ \frac{d\ln X_{ij}}{d\ln \tau_{kl}}&=(1-\sigma)\,1\{ij=kl\}+(1-\sigma)\frac{d\ln p_i}{d\ln \tau_{kl}}+(\sigma-1)\frac{d\ln P_j}{d\ln \tau_{kl}}+\frac{d\ln Y_j}{d\ln \tau_{kl}}\end{aligned}$$

A.6.3 Elasticity of Welfare to Expected Trade Cost without Labor Mobility

The model becomes the standard Armington model when there is no labor mobility, we can apply Allen et al. (2020a)'s approach to first get

$$\begin{aligned}\frac{d\ln p_l}{d\ln \tau_{ij}}&=-\phi X_{ij}\left(b_{l,i}+b_{l,N+j}\right)\\ \frac{d\ln P_l}{d\ln \tau_{ij}}&=-\phi X_{ij}\left(b_{N+l,i}+b_{N+l,N+j}\right)\end{aligned}$$

where p_s is normalized to 1, $\phi\equiv\sigma-1$, $\psi\equiv-\frac{\alpha+1}{\alpha+\beta}$, and b_{ij} is the $i-jth$ element of

$$A^{-1}\equiv\begin{bmatrix}(1+\psi)\,diag\,(Y)-(1+\psi)\,X+\phi diag\,(Y) & -\psi diag\,(Y)-(\phi-\psi)\,X\\ \phi X' & -diag\,(Y)\end{bmatrix}^{-1}$$

Then, for the welfare

$$W_i = \frac{w_i}{P_i} = \frac{p_i A_i}{P_i}$$
$$\Rightarrow \frac{d \ln W_i}{d \ln \tau_{kl}} = \frac{d \ln p_i}{d \ln \tau_{kl}} - \frac{d \ln P_i}{d \ln \tau_{kl}}$$

and

$$\begin{aligned}
\frac{d \ln L_i}{d \ln \tau_{kl}} &= 0 \\
\frac{d \ln w_i}{d \ln \tau_{kl}} &= \frac{d \ln p_i}{d \ln \tau_{kl}} \\
\frac{d \ln Y_i}{d \ln \tau_{kl}} &= \frac{d \ln w_i}{d \ln \tau_{kl}} \\
\frac{d \ln X_{ij}}{d \ln \tau_{kl}} &= (1-\sigma) 1\{ij = kl\} + (1-\sigma) \frac{d \ln p_i}{d \ln \tau_{kl}} + (\sigma - 1) \frac{d \ln P_j}{d \ln \tau_{kl}} + \frac{d \ln Y_j}{d \ln \tau_{kl}}
\end{aligned}$$

Recall that in the calculation of $d \ln W_i / d \ln \tau_{kl}$ above, we normalize $w_s = 1$ when using Allen and Arkolakis (2019)'s formula, but normalize $p_s = 1$ when using Allen et al. (2020a)'s formula. As $w_s = p_s / A_s$, and there is no labor mobility as well as externality now, these two versions of normalization are equivalent.

A.6.4 Elasticity of Trade Flow to Direct Trade Cost

Recall that $\Xi_{kl} = \sum_i \sum_j X_{ij} \left[\tau_{ij} / \left(\tau_{ik} t_{kl} \tau_{lj} \right) \right]^{\theta}$. The elasticity of Ξ_{kl} with respect to t_{kl} is

$$\begin{aligned}
\frac{d \ln \Xi_{ij}}{d \ln t_{kl}} &= \frac{1}{\sum_{m'n'} X_{m'n'} \pi^{ij}_{m'n'}} \sum_{mn} \left(\frac{d X_{mn}}{d \ln t_{kl}} \pi^{ij}_{mn} + X_{mn} \frac{d \pi^{ij}_{mn}}{d \ln t_{kl}} \right) \\
&= \frac{1}{\sum_{m'n'} X_{m'n'} \pi^{ij}_{m'n'}} \sum_{mn} X_{mn} \pi^{ij}_{mn} \left[\sum_{k'l'} \frac{d \ln X_{mn}}{d \ln \tau_{k'l'}} + \theta \left(\pi^{kl}_{mn} - \pi^{kl}_{mi} - \pi^{kl}_{jn} - 1\{ij = kl\} \right) \right]
\end{aligned}$$

where $d X_{mn} / d \ln \tau_{k'l'}$ is obtained in Section A.6.1 and $\pi^{kl}_{mn} \equiv d \ln \tau_{mn} / d \ln t_{kl}$ is provided by (29).