DISCUSSION PAPER SERIES

DP15905

TO BE CONNECTED OR NOT TO BE CONNECTED? THE ROLE OF LONG-HAUL ECONOMIES

Hans Koster, Takatoshi Tabuchi and Jacques-François Thisse

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

PUBLIC ECONOMICS

ISSN 0265-8003

TO BE CONNECTED OR NOT TO BE CONNECTED? THE ROLE OF LONG-HAUL ECONOMIES

Hans Koster, Takatoshi Tabuchi and Jacques-François Thisse

Discussion Paper DP15905
Published 11 March 2021
Submitted 09 March 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

TO BE CONNECTED OR NOT TO BE CONNECTED? THE ROLE OF LONG-HAUL ECONOMIES

Abstract

We investigate what regions are the winners and losers from large transportation infrastructure improvements. We argue that long-haul economies - implying that the marginal transportation cost decreases with distance - play a pivotal role in understanding the location choices of firms. Using data from Japan and the Netherlands, we first establish that long-haul economies are an important feature of modern transportation networks. Then, we develop a simple model to show that improvements in transportation infrastructure have non-trivial impacts on the location choices of firms. While these investments are often beneficial to large regions, they may be detrimental to small intermediate regions, implying job losses. Using data on Japan's Shinkansen, we confirm that 'in-between' municipalities that are connected to the HSR witness a sizable decrease in employment.

JEL Classification: H40, O18, R30, R42

Keywords: accessibility, Transport Infrastructure, long-haul economies, Regional Development

Hans Koster - h.koster@vu.nl
Vrije Universiteit Amsterdam and CEPR

Takatoshi Tabuchi - takatoshi.tabuchi@gmail.com
Chuo University

Jacques-François Thisse - jacques.thisse@uclouvain.be
Higher School of Economics (St Petersburg) and CEPR

Acknowledgements
We are grateful to S. Proost and M. Turner for useful comments and discussions. We also thank participants of a seminar at the MUFG Bank. The first and third authors acknowledge the support of the HSE University Basic Research Program.

To be connected or not to be connected? The role of long-haul economies*

Hans R.A. Koster† Takatoshi Tabuchi‡ Jacques-François Thisse§

March 9, 2021

Abstract

We investigate what regions are the winners and losers from large transportation infrastructure improvements. We argue that long-haul economies – implying that the marginal transportation cost decreases with distance – play a pivotal role in understanding the location choices of firms. Using data from Japan and the Netherlands, we first establish that long-haul economies are an important feature of modern transportation networks. Then, we develop a simple model to show that improvements in transportation infrastructure have non-trivial impacts on the location choices of firms. While these investments are often beneficial to large regions, they may be detrimental to small intermediate regions, implying job losses. Using data on Japan's Shinkansen, we confirm that 'in-between' municipalities that are connected to the HSR witness a sizable decrease in employment.

Keywords: long-haul-economies; firm location; high-speed railway; hub effect

JEL Classification: D43; R12; R40

*We are grateful to S. Proost and M. Turner for useful comments and discussions. We also thank participants of a seminar at the MUFG Bank. The first and third authors acknowledge the support of the HSE University Basic Research Program.

†Department of Spatial Economics, Vrije Universiteit Amsterdam, HSE University (Russia), Tinbergen Institute, and CEPR. E-mail: h.koster@vu.nl.

‡Faculty of Global Management, Chuo University (Japan). E-mail: ttabuchi@tamacc.chuo-u.ac.jp.

§Institute of Developing Economies (Japan), HSE University (Russia) and CEPR. E-mail: jacques.thisse@uclouvain.be.

1 Introduction

There seems to be a general agreement that the supply of transportation infrastructure has a positive effect on the development of the so-connected regions (see the survey by Redding and Turner, 2015). For example, Donaldson and Hornbeck (2016) and Donaldson (2018) document the positive effect of railroads in the U.S. from 1870 to 1890 and in colonial India from 1870 to 1930, respectively. Using recent British micro data at a very detailed geographical scale, Gibbons *et al.* (2019) find that road network improvements have a significant and positive impact on the number of new establishments and the output per worker in existing establishments, while Aggarwal (2018) finds that paved roads connecting rural areas in contemporary India generate substantial welfare gains for the rural population. Likewise, Ahlfeldt and Feddersen (2018) provide evidence that access to a high-speed railway connecting Cologne and Frankfurt (Germany) leads to an increase in GDP by 8.5% in three counties with intermediate stops. Hence, it is no surprise that transport investments have been, and still are, one of the main instruments used by governments and international institutions to boost regional development and to reduce regional disparities by improving accessibility of lagging regions.

Large amounts of public money are thus invested in the development of transportation infrastructure. To illustrate, in 2018 Japan spent about 1% of its GDP on infrastructure construction and maintenance, while China spent a staggering 5.5% (De Best, 2020). In the late 1990s, the European Commission launched a large transport infrastructure program with 30 priority projects for a total amount of about 600 billion euros. However, an ex ante assessment of this package yields interesting findings (Proost *et al.*, 2014). First, only 12 out of the 22 projects considered pass a simple cost-benefit analysis test. Second, the projects do not necessarily favor the poorest regions and so do not seem to decrease regional disparities within the EU. In the U.S., the federal government finances a large share of interstate highways by using revenues from the gasoline tax. The econometric evidence provided by Baum-Snow (2007) and Duranton and Turner (2012) suggests that highways tend to be disproportionately allocated to smaller metropolitan areas because roads are more expensive to build in larger urban areas. Because the latter areas get fewer roads per capita, urban growth then falls behind smaller urban areas. Likewise, the works of Faber (2014) and Baum-Snow *et al.* (2017) suggest that the opening of new transportation links involves winners and losers. More specifically, they show that the construction of highways in China is mainly beneficial to the large metropolitan areas that are connected by these new links, while the in-between regions may be hurt. In the same vein, Qin (2013) finds that non-urban counties on upgraded Chinese railway lines experienced reductions in GDP per capita following the upgrade. Also, Asher and Novosad (2020) find that, four years after construction, India's $40 billion national rural road construction program did not affect much agricultural outcomes. In particular, better market connections did not make remote areas much better off. Hence, the literature on the effects of large transportation infrastructure investments is rather mixed (Blanquart and Koning, 2017).

The purpose of this paper is to shed new light on the issue of what regions benefit or lose from transportation investments, by providing a reconciliation between the above two strands of literature. We accomplish this by showing that new links and technologies affect transportation costs in ways that are less trivial than what the almost universal, but somewhat too naive, use of iceberg transportation cost suggests. Indeed, it went largely unnoticed in the trade and geography literature that the way in which (network) distance affects transportation costs has a fundamental impact on firms' location choices. For centuries, transportation costs were mainly determined by the distance covered and the marginal cost increased sharply with distance. Such a situation describes quite well periods in which moving commodities and people was both very costly and risky. Thanks to technological and organizational advances, ships were to become able to cross longer distances in one go. This revolution in maritime transportation permitted the European expansion toward Africa and Asia in the 15^{th} to 16^{th} centuries (Brook, 2008; Subrahmanyam, 2012). On land, only since the advent of railways substantial progress occurred at the time of the Industrial Revolution, which in turn allows for the access to mass markets within Europe and the United States (Bogart, 2014). In both cases, long-distance journeys and shipments became relatively less expensive. As investments in infrastructure and equipment grew further, distance mattered less. Under these circumstances, the marginal transportation cost no longer increases but decreases with distance. Transporting commodities or people is then characterized by what transportation economists call *long-haul economies* (Boyer, 1997; Quinet and Vickerman, 2004).

By reducing the value of transportation costs between distant regions, but less between close regions, long-haul economies (LHE) reshuffle the relative position of locations within transportation networks, and thus affect the location of production in a complex manner. They are different from a simple reduction in shipping costs. Our study aims to investigate how LHE affect firms' location choices. To illustrate the forces at work, consider a firm that buys an input in one market and sells its output in another; both markets are connected by a link. The optimal location of the firm, which minimizes the sum of transportation costs, can be viewed as the equilibrium point of a system of two forces generated by the need for proximity to the output and input markets. When the transportation cost function is strictly convex, the intensity of these forces increases rapidly with distance. As a result, the system of forces is in equilibrium when the firm chooses the location where the two marginal transportation costs are equal. In this case, the cost-minimizing location is typically situated in between the two markets. By contrast, when the transportation cost function is strictly concave in distance, the marginal costs are never equal, and thus the firm chooses to set up at the market that has the lower marginal cost. Hence, the firm's optimal locations qualitatively differ under LHE or long-haul *diseconomies*.

We extend this idea by showing how modern means of transportation and the LHE they bring about affect the pattern of production in ways that are not a priori expected. To make our main point transparent, we consider toy-networks, such as linear or triangular spaces. We also use a simple location model to study how LHE interact with the trade-off between a 'hub effect' and a 'home-market effect'. In this respect, we have chosen to isolate the effects of LHE on firms' locations from endogenous market or industry size considerations. However, we will argue that allowing for migration should not reverse our main conclusions. Furthermore, we provide a detailed empirical analysis that confirms our predictions on the impact of new transportation links on employment shares in regions that are connected to the Shinkansen network in Japan. Since new infrastructure, such as a highway or high-speed rail (HSR), is often the source of LHE, our results shed light on an issue that has been, and still is, at the center of debates among scientists and policy-markets, i.e., *does a region attract firms when it is connected to a new transportation infrastructure?*

Our main findings are organized around two main ideas. First, we show that *the construction of a new link connecting two large regions may be detrimental to the intermediate regions it passes through* (Proposition 1). In the absence of LHE, an intermediate region may host a more than proportionate share of firms when its average distance to all markets is minimized, a result called the *hub effect* (Krugman, 1993; Behrens *et al.*, 2007; Mori, 2012). We show that this effect is undermined by the presence of LHE, i.e., fewer firms choose to set up in the intermediate region as the intensity of LHE increases because of a so-called *home market effect.* If the intermediate region is connected to the new infrastructure, so that shipping goods becomes cheaper, the two large regions become relatively more attractive because the home market effect gets stronger. Hence, due to LHE, *an intermediate region may, but need not, attract more firms when it is connected to a new transportation link.* This ambiguous finding may explain why the construction of a highway ramp or an HSR station does not necessarily deliver its sought-after payoffs.

Second, when a new and rapid transportation link between two metropolitan areas is established, the highway or the HSR often passes through several intermediate regions without making a stop. In this case, it is commonplace for local governments and regional interest groups to lobby at the federal/national government and international bodies such as the European Commission or the World Bank for their region to be connected to the new infrastructure. Yet, given what we have seen above, it is not clear whether being connected is better or worse than remaining unconnected. This is the next question we investigate in the special case of a network formed by two endpoints and two intermediate regions where one of them is connected to the new link whereas the other is not. To put it differently, shipping goods to and from one of the two intermediate regions benefit from LHE while the other does not, which implies higher transportation costs. Once more, the answer is ambiguous: *being connected is not sufficient for an intermediate region to retain more firms than the unconnected one* (Proposition 2). More specifically, when its market is small, the connected region does not benefit from a better connection to the large regions because the home market effect mitigates the hub effect. However, when the intermediate regions are sufficiently large, being connected is always an advantage in that the connected region has a higher firm share than the unconnected one. In this case, the unconnected region suffers from being more separated from the two large metropolitan regions, while the hub effect overcomes the home market effect.

The above findings may look like exotica that have no policy or practical relevance. This is why we undertake a reduced-form empirical analysis to test their meaning by investigating *the effects on employment of being connected to the Shinkansen network in Japan.* There are several reasons why studying the Shinkansen (which means 'new trunk line') is important. First, one of the main objectives of the Shinkansen was to promote economic growth and development outside Tokyo (Sato, 2015). Second, we will show that the Shinkansen displays strong LHE. Third, out of 160 million of passengers per year, a very large share (i.e., about 65% in 2010) are technical workers and business travellers. Such a high number strongly suggests that the Shinkansen may be considered as a transportation mode that affects significantly firms' location choices through the travel of non-production workers whose share in Japan has increased from 22% to 41% between 1952 and 2015. Last, the first Shinkansen lines were built more than 50 years ago, so that one may expect their long run effects to have materialized. All of this makes the Shinkansen, a natural candidate to study the impact of LHE on the location of firms.

Since a transportation improvement not only affects the connected regions, but the whole array of regions through their relative position within the network, we consider Japan's main islands, i.e., Hokkaido, Honshu, Shikoku, and Kyushu, which account for 99% of the Japanese population. Our sample thus contains $1,658$ municipalities.[1] To explicitly test Proposition 2, and to mitigate endogeneity issues, we exclude central municipalities and only keep in-between municipalities. Furthermore, following common practice in the literature, we construct instruments based on plans for the Shinkansen designed in 1942 and 1972. We first regress the log of the change in employment shares between 1957 and 2014 on whether the municipality has a Shinkansen station in 2014. Different specifications deliver a consistent picture: *areas lose firms when they are connected to the Shinkansen.* The effects ranges from about 10 to 40%. While this effect may seem large, it is very much in the same order of magnitude as Faber (2014) and Baum-Snow *et al.* (2017) who follow a related approach.

A last remark is in order. Our empirical section focuses on HSR. It is nevertheless noteworthy that the theoretical results of Section 4 hold true for any other transportation model that displays

[1] We also show the robustness of our results by using alternative geographical areas, such as commuting zones.

long-haul economies, e.g., maritime or air transport over long distances (Mori, 2012; Campate and Yanagizawa-Drott, 2018).

The remainder of the paper is organized as follows. In Section 2, we provide empirical evidence that LHE exist for travel times and network distances in Japan and the Netherlands. The model is presented in Section 3. In Section 4, we study the impact of LHE on the relative (dis)advantage for a region to be connected when others are unconnected to a new transportation link. Section 5 provides empirical support of the results obtained in the previous section. Section 6 concludes.

2 Empirical evidence for long-haul economies

Using data on travel times and network distances from Japan and the Netherlands, we provide here evidence that LHE are a defining characteristic of modern transportation modes. For details on the data we refer to Appendix A. We adopt a simple framework, which is in line with Couture *et al.* (2018), to investigate the importance of LHE. That is, we regress (free-flow) travel time t_{ij} between origin i and destination j on network distance d_{ij}:[2]

$$\log t_{ij} = (1-\eta)\log d_{ij} + \mu_i + \mu_j + \epsilon_{ij},$$

where $\eta \in [0,1)$ indicates the presence of LHE; μ_i and μ_j are origin and destination fixed effects that control for measurement error in travel times; ϵ_{ij} is an unobserved component of travel times that is assumed to be uncorrelated to d_{ij}.

Table 1 reports the results. In column (1) we focus on roads in Japan. We find that $\hat{\eta} = 0.094$, which is substantially and statistically significantly higher than 0. Hence, when distance increases by 1%, travel times increase by only 0.9%. LHE arise because speeds are higher on highways. These results are in accordance with Couture *et al.* (2018), who also find that for U.S. cities, time costs per kilometer go down once the distance increases. We find even stronger evidence for LHE in the railway network in Japan, where $\hat{\eta} = 0.327$, because the Shinkansen considerably shortens travel times on longer distances.

In columns (3) and (4) of Table 1, we consider a smaller country, i.e., the Netherlands. In column (3), we find again strong evidence for LHE in road travel, as $\hat{\eta} = 0.206$. A reason why LHE are stronger than in Japan's road network may be that we have very detailed information on local streets, where speeds are considerably lower. For railway travel time, we find $\hat{\eta} = 0.138$. Although we still find strong and statistically significant evidence for LHE, these are less pronounced compared to Japan because there are hardly high-speed trains in the Netherlands.[3]

[2] We ignore issues of schedule delay in calculating travel times, because we do not have information on congestion and peoples' preferred arrival times.

[3] More specifically, the first HSR has been opened in 2009 connecting Amsterdam, Rotterdam and Breda. On

Table 1 — Regression results
(Dependent variable: the log of travel time)

	Japan		*The Netherlands*	
	(1)	(2)	(3)	(4)
	Roads	Railways	Roads	Railways
Long-haul economies, $\hat{\eta}$	0.0942***	0.3274***	0.2056***	0.1384***
	(0.0028)	(0.0062)	(0.0009)	(0.0059)
Origin fixed effects	Yes	Yes	Yes	Yes
Destination fixed effects	Yes	Yes	Yes	Yes
Number of observations	3,059,706	2,958,442	16,180,503	144,776
R^2	0.9860	0.9449	0.9838	0.9304

Notes: Origins and destinations refer to municipalities (Japan), postcodes (for the road network in the Netherlands) or stations (for the railway network in the Netherlands). We cluster standard errors at the origin and destination. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

3 The model

The economy has two sectors and $M \geq 2$ regions. One sector produces a continuum of varieties of a horizontally differentiated product under monopolistic competition and increasing returns, using labor. Firms choose where to produce across the M regions. The other sector produces a homogeneous good under perfect competition and constant returns, using labor. Region j is populated with $L_j \equiv \theta_j L$ consumers/workers, with $\Sigma_{j=1}^{M} \theta_j = 1$. In what follows, we choose the unit of labor for the total population L to be equal to 1, so that the size of region j is given by θ_j. Consumers are geographically immobile, but perfectly mobile between sectors.

The utility of a consumer living in region j is given by

$$U_j = H_j^{1-\mu} Q_j^{\mu}, \quad 0 < \mu < 1 \tag{1}$$

where

$$Q_j = \left[\sum_{i=1}^{M} \left(\int_{\Omega_i} q_{ij}(\omega)^{\frac{\sigma-1}{\sigma}} \mathrm{d}\omega \right) \right]^{\frac{\sigma}{\sigma-1}}, \tag{2}$$

where H_j is the consumption of the homogeneous good, $q_{ij}(\omega)$ is the individual consumption in region j of variety $\omega \in \Omega_i$, where Ω_i is the set of varieties produced in region i, and $\sigma > 1$ the elasticity of substitution between any two varieties.

Because of symmetry, we may drop the variety label ω. Then, the utility maximization of (1) subject to the income constraint yields the following individual demand in region j for a variety

this line, trains travel over 200km/h. The maximum speed of regular 'inter-city trains' is just 140km/h.

produced in region i:

$$q_{ij} = \frac{p_{ij}^{-\sigma}}{P_j^{1-\sigma}} w_j, \tag{3}$$

where p_{ij} is the price of a variety produced in region i and sold in region j, w_j the wage in region j, while

$$P_j = \left(\sum_{k=1}^{M} n_k p_{kj}^{1-\sigma} \right)^{\frac{1}{(1-\sigma)}} \tag{4}$$

is the price index in this region and n_k is the mass of firms in region k.

After normalization, one unit of the homogeneous good requires one unit of labor. This good can be traded costlessly across regions and is chosen as the numéraire. Hence, wages are equalized to one across regions and sectors ($w_i = 1$). Turning to the differentiated good sector, producing of a variety requires a fixed and a marginal labor requirements, $F > 0$ and $c > 0$.[4] Shipping one unit of a variety from i to j is given by the iceberg transportation cost $\tau_{ij} > 1$, which means that τ_{ij} units of the variety must be shipped for one unit to arrive at destination. We will see in the next sections how this modeling trick proposed long ago by Samuelson (1954) may be extended to cope with LHE.

A firm established in region i maximizes its profit given by

$$\Pi_i = \sum_{j=1}^{M} (p_{ij} - c\tau_{ij}) L_j q_{ij} - F. \tag{5}$$

Profit maximization with respect to p_{ij} yields the equilibrium price

$$p_{ij}^* = \frac{\sigma}{\sigma - 1} c\tau_{ij}. \tag{6}$$

Free entry and exit imply that profits are non-positive in equilibrium. It then follows from (5) and (6) that firms' equilibrium output in region i must satisfy:

$$\sum_{j=1}^{M} \tau_{ij} L_j q_{ij} \leq \frac{F(\sigma - 1)}{c}. \tag{7}$$

Hence, all firms have the same size, which is equal to

$$Q = \frac{F(\sigma - 1)}{c}.$$

[4] The model can easily be extended to account for region-pair-specific total factor productivity c_{ij}.

Let

$$\phi_{ij} \equiv \tau_{ij}^{1-\sigma} \in (0,1),$$

which has the nature of a *transportability rate*: the higher is ϕ_{ij}, the less costly it is to ship a variety between i and j. Plugging (3) and (4) into (7), multiplying both sides by p_{ii}, and using (6), we get:

$$\frac{1}{\sigma F}\sum_{j=1}^{M}\frac{\phi_{ij}L_j}{\sum_k \phi_{kj}n_k} \leq 1, \tag{8}$$

where the equality prevails when $n_i^* > 0$.

Following Behrens *et al.* (2007), the equilibrium conditions (8) can be rewritten as follows:

$$\frac{1}{\sigma F}\Phi\,\mathrm{diag}(\Phi\mathbf{n})^{-1}\mathbf{L} \leq \mathbf{1},$$

where

$$\Phi \equiv \begin{pmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1M} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{M1} & \phi_{M2} & \cdots & \phi_{MM} \end{pmatrix} \qquad \mathbf{n} \equiv \mathrm{diag}\begin{pmatrix} n_1 \\ \vdots \\ n_M \end{pmatrix} \qquad \mathbf{L} \equiv \mathrm{diag}\begin{pmatrix} L_1 \\ \vdots \\ L_M \end{pmatrix} \qquad \mathbf{1} \equiv \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

If an interior equilibrium exists ($n_i^* > 0$), it is given by

$$\mathbf{n}^* = \frac{\mu}{\sigma F}\Phi^{-1}\mathrm{diag}\left(\Phi^{-1}\mathbf{1}\right)^{-1}\mathbf{L}. \tag{9}$$

Hence, the *firm share* in region i is given by

$$\lambda_i^* \equiv \frac{n_i^*}{N}, \qquad i = 1, ..., M, \tag{10}$$

where

$$N \equiv \sum_{i=1}^{M} n_i^* \equiv \frac{L}{\sigma F}$$

is the total number of firms, which is therefore given. Thus, a higher value of λ_i^* means a higher employment in region i. Importantly, like most economic geography models based on the CES, the mass of firms N and the total output NQ are constants (see, e.g., Baldwin *et al.*, 2003). Hence, in our model, new links do not lead to new activities but displace them from some regions to others. The creation of new activities would require another setting (see, e.g., Monte *et al.*, 2018, and Hayakawa *et al.*, 2021).

Since N is constant, we can work with the firm shares λ_i^* only. The share λ_i^* of firms operating in region i is endogenous and depends on the parameters of the economy. When ϕ_{ij} changes, profits rise in some regions and fall in others. In the former, new firms are launched; in the latter, a few incumbents shut down.

Finally, wage equalization requires the production of the homogeneous good in all regions. For this condition to be satisfied, the total mass of consumers/workers in each region must exceed the total labor requirement needed to produce the differentiated sector. That is, $L_i > n_i (cQ + F)$ must hold for all i. This is equivalent to $\theta_i > \mu\lambda_i^*$. We thus assume:

$$\mu < \theta_i, \qquad i = 1, \ldots, M. \tag{11}$$

Hence, a region cannot be too small to accommodate firms.

The hub effect. Consider a 3-region economy where the size of region i is $\theta_i > 0$ for $i = 1, 2, 3$. We assume that regions 1 and and 3 have the same size, i.e., $\theta_2 = \theta < 1/3$ so that $\theta_1 = \theta_3 = (1 - \theta)/2 > 1/3$. The three regions are connected by a one-dimensional network. The regions 1 and 3 are located at the endpoints of the link, while the intermediate region 2 is equally distant from 1 and 3: $d_{12} + d_{23} = d_{13}$.

The trade and geography literature typically assumes that shipping varieties directly between regions 1 to 3 is as costly as shipping them via the in-between region 2, that is, $\tau_{13} = \tau_{12}\tau_{23} = \tau^2 > 1$ and $\phi_{13} = \phi_{12}\phi_{23} = \phi^2$ where $\phi = \tau^{1-\sigma} < 1$. The corresponding transportation cost matrix is then as follows:

$$\Phi = \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix}. \tag{12}$$

Thus, the equilibrium production pattern is the outcome of the trade-off between a centrifugal force and a centripetal force: the former is a home market effect in which each large region ($i = 1$ and $i = 3$) attracts firms from the small region, while the latter is a hub effect in which the centrality of region 2 allows firms located therein to supply regions 1 and 3 at the lowest total transportation costs.

Since $\theta_1 = \theta_3$, the symmetry of the market space suggests that regions 1 and 3 accommodate the same share of firms, which is equal to $(1 - \lambda_2^*)/2$. As a result, determining the spatial equilibrium is equivalent to finding the equilibrium value of λ_2. Using (9), (10) and (12) implies that the equilibrium share of firms established in region 2 is given by

$$\lambda_2^* = \frac{\theta - \phi}{1 - \phi} + \frac{2\phi\theta}{(1 - \phi)^2}, \tag{13}$$

which is positive if and only if region 2 has a *minimum size* given by $\underline{\theta} \equiv \phi(1-\phi)/(1+\phi) > 0$. In other words, a locational advantage is not sufficient to attract firms: the corresponding region must have a sufficiently large market. Differentiating $\underline{\theta}$ with respect to ϕ shows that $\underline{\theta}$ increases with ϕ at a decreasing rate. Indeed, when shipping the good is cheaper, the two large regions become relatively more attractive because the home market effect gets stronger, which implies that the intermediate region must be larger to attract firms.

Since

$$\frac{d\lambda_2^*}{d\phi} = \frac{3\theta + \phi + \theta\phi - 1}{(1-\phi)^3},$$

the firm share in region 2 first decreases until $\phi = (1-3\theta)/(1+\theta)$ and, then, increases. In other words, *a steady decrease of transportation costs* ($\phi \uparrow$) *first leads to the gradual dispersion of firms through the a growing mass of firms in the two large regions*, due to the home market effect *and, then, to the gradual agglomeration of firms in the in-between region*, due to the hub effect. This contradicts the prediction by the two-region model that the agglomeration process is monotone (see, e.g., Baldwin *et al.*, 2003). Once more, the two-region setting leads to results that no longer hold when mildly more general geographies are considered.

Let

$$\bar{\theta} \equiv \frac{1-\phi}{3-\phi} \in (0, 1/3).$$

be the solution to $\lambda_2^* = \theta$. Although region 2 is smaller than region 1 or 3 ($\bar{\theta} < \theta < 1/3 < \theta_1 = \theta_3$, the mass of firms located in region 2 is larger than the mass of firms in region 1 or 3 when $\theta > \bar{\theta}$ ($\lambda_2^* > \lambda_1^* = \lambda_3^*$). In this case, the locational advantage of region 2 overcomes its market size disadvantage.

Intermediate inputs. The above setting is isomorphic to a model in which the final sector produces a homogeneous good under perfect competition and constant returns by using differentiated intermediate goods (Ethier, 1982). The share of the final sector in region i is given by L_i. The production function of the final sector is given by (2) where $q_{ij}(\omega)$ is the quantity of input ■ produced in region j and used in region i. The intermediate goods are produced under monopolistic competition and increasing returns by using labor. An input may be a physical good shipped to a firm belonging to the final sector or a task performed by a technical worker who travels to such a firm. The aim of the model is then to determine the interregional distribution λ_i of intermediate firms. Since its focuses on inputs, the intermediate variety model fits better our econometric analysis of the impact of the Shinkansen than the final variety model. Since the latter is more widely discussed than the former, we have chosen to use the final variety model to present our theoretical developments. However, the reader should keep in mind that our results also hold true for business-to-business service-providers.

4 Is being connected a (dis)advantage?

4.1 Long-haul economies

We define what LHE are and how to model them within the framework of the iceberg transportation cost. Let X be a (finite or continuous) location set and $d(i,j)$ the (shortest) distance between locations i and j. Consider an iceberg transportation cost function $\tau(d)$ which is continuous, strictly increasing in distance d and such that $\tau(0) = 1$. As seen in the foregoing, the transportation cost function $\tau(d)$ used in the trade and geography literature satisfies what we may call the 'multiplicativity property': $\tau(d(i,k)) = \tau(d(i,j))\,\tau(d(j,k))$ when $d(i,j)+d(j,k) = d(i,k)$ for $i \neq j \neq k$. The only continuous function that satisfies this condition is $\tau(d) = \exp(\kappa d)$ where $\kappa > 0$ is a spatial friction parameter. Hence, the multiplicativity property implies that the transportation cost is strictly *convex* in d, i.e., there are long-haul *dis*economies (McCann, 2005). Since the triangular inequality $\tau(d(i,k)) < \tau(d(i,j)) + \tau(d(j,k))$ does not hold, transshipment via region j is less costly than direct shipment from regions i to k, which seems counterintuitive.

Assume now that the transportation cost function $\tau(d)$ displays LHE or, equivalently, that $\tau(d)$ is strictly *concave* in d. Therefore, $\log\tau(d)$ is also strictly concave in d, so that $(1-\sigma)\log\tau(d)$ is strictly convex in d. Since $d(i,j) + d(j,k) = d(i,k)$, we have:

$$(1-\sigma)\log\tau(d(i,j)) + (1-\sigma)\log\tau(d(j,k)) < (1-\sigma)\log\tau(d(i,k)).$$

As the transportability rate is given by $\phi(\tau(d)) \equiv (\tau(d))^{1-\sigma}$, the above expression is equivalent to

$$\phi(\tau(d(i,j)))\;\phi(\tau(d(j,k))) < \phi(\tau(d(i,k))). \tag{14}$$

Setting $\phi \equiv \phi(\tau(d(i,j))) = \phi(\tau(d(j,k))) < 1$ and $\phi_{ik} \equiv \phi(\tau(d(i,k))) < \phi$, there exists $\eta \in [0,1)$ such that $\phi^{2(1-\eta)} = \phi_{ik}$ where the value $\eta = 1$ is excluded because it amounts to assuming zero transportation costs. It then follows from (14) that LHE prevail if and only if $\phi^2 < \phi^{2(1-\eta)}$. Since $\phi < 1$, this inequality holds if and only if $0 < \eta < 1$, which agrees with the empirical evidence provided in Section 2. This shows that iceberg transportation costs are not incompatible with the presence of LHE. Note also that the strict concavity of $\tau(d)$ implies the triangular inequality: $\tau(d(i,k)) < \tau(d(i,j)) + \tau(d(j,k))$. Since $\phi^{1-\eta}$ increases at a decreasing rate with the transportability of the differentiated good, η may be interpreted as the *elasticity of the LHE*: the higher η, the higher $\phi^{1-\eta}$, that is, the lower the cost of shipping varieties between i and k. When $\eta > 0$, the direct transfer between regions i and k is less expensive than the transfer through the in-between region j. For example, this arises when i and k are connected by a highway or an

HSR that by-passes j.

4.2 The three-region case

We first discuss the effects of LHE in the 3-region setting discussed in Section 3. More specifically, we assume that a new transportation infrastructure, such as a highway or an HSR, connects the large regions 1 and 3 whereas region 2 is unconnected to this new link. The transportation matrix (12) thus becomes:

$$\Phi = \begin{pmatrix} 1 & \phi & \phi^{2(1-\eta)} \\ \phi & 1 & \phi \\ \phi^{2(1-\eta)} & \phi & 1 \end{pmatrix}. \tag{15}$$

Using (9), (10) and (15), it is readily verified that the equilibrium share of firms established in region 2 is given by

$$\lambda_2^* = \frac{\theta - \phi}{1 - \phi} + \frac{2\phi\theta}{1 - 2\phi + \phi^{2(1-\eta)}}. \tag{16}$$

Less expected, we have:

$$\frac{d\lambda_2^*}{d\eta} = \frac{4\phi^{3+2\eta}\theta \log \phi}{\left[(1-2\phi)\,\phi^{2\eta} + \phi^2\right]^2},$$

which is negative because $\log \phi < 0$ while the bracketed term is positive. Thus, stronger LHE ($\eta \uparrow$) decrease λ_2^* and increase λ_1^* and λ_3^*. In other words, the construction of a new and efficient transportation infrastructure that connects the large regions weakens the hub effect.

The following proposition comprises a summary.

Proposition 1 — The equilibrium mass of firms located in the intermediate region decreases with the elasticity of LHE ($\eta \uparrow$).

Assume now that region 2 is connected to the new link, that is, $\phi_{12} = \phi_{23} = \phi^{1-\eta} < \phi$. In this case, the equilibrium mass of firms established in region 2 is given by

$$\lambda_2^{\circledast} = \frac{\theta - \phi^{1-\eta}}{1 - \phi^{1-\eta}} + \frac{2\phi^{1-\eta}\theta}{\left(1 - \phi^{1-\eta}\right)^2},$$

where the superscript $\circledast$ refers to the equilibrium share of firms when region 2 is connected.

The following question then arises naturally: does connecting the intermediate region to the new infrastructure decrease or increase the local employment level? Or, more formally, is $\lambda_2^{\circledast}$ smaller or larger than λ_2^* given by (16)? Differentiating $\lambda_2^{\circledast}$ with respect to η yields

$$\frac{d\lambda_2^{\circledast}}{d\eta} = \phi^{1+\eta} \ln \phi \frac{(1-3\theta)\phi^{\eta} - (1+\theta)\,\phi}{\left(\phi^{\eta} - \phi\right)^3},$$

which is negative when $\theta > (1 - \phi^{1-\eta})/(3 + \phi^{1-\eta})$ and nonnegative otherwise. Therefore, we cannot answer the above questions by 'yes' or by 'no'. Instead, two cases may arise.

On the one hand, when the local market is small, fewer firms set up in the intermediate region when this region is connected to the new link. Furthermore, stronger LHE render region 2 less attractive. Indeed, as η increases, firms gradually leave region 2 to benefit from being established in a large market without losing much for being away from a small market. In this case, the hub effect is offset by the home market effect generated by the large regions. In the limit, a new transportation link that gives rise to very strong LHE may even cause the hollowing out of region 2.

On the other hand, when the local market is sufficiently large, an intermediate region hosts more firms when it is connected to the link. Indeed, firms now enjoy the benefit of a large local market while having a good access to the big markets. As a result, region 2 gains firms at the expense of regions 1 and 3. In this case, the construction of a new transportation link that connects the three regions is detrimental to the large ones and beneficial to the intermediate one.

To sum up, *whether an in-between region benefits from being connected to a new transportation infrastructure depends on its size*: the larger it is, the more likely this region will benefit from the connection.

4.3 The four-region case

We now turn our attention to an economy with four regions, which are located equidistantly and connected by a linear network. Regions 1 and 4 are located at the endpoints of the link and host a larger population $\theta_1 = \theta_4 = (1 - \theta)/2 > 1/4$ than regions 2 and 3 with $\theta_2 = \theta_3 = \theta < 1/4$. A new transportation infrastructure connects the large regions 1 and 4 whereas regions 2 and 3 are unconnected to this new link. The transportation cost matrix is then as follows:

$$\Phi = \begin{pmatrix} 1 & \phi & \phi^2 & \phi^{3(1-\eta)} \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^{3(1-\eta)} & \phi^2 & \phi & 1 \end{pmatrix}.$$

In this case, the equilibrium firm shares are given by

$$\lambda_2^* = \lambda_3^* = \frac{(2\theta - \phi)\,\phi^3 + \phi^{3\eta}\left[2\theta\left(1 - \phi^2 - \phi^3\right) - \left(1 - \phi + \phi^2\right)\right]}{2\,(1 - \phi)\,\phi^3 + 2\phi^{3\eta}\left(1 - 2\phi + \phi^3\right)}.$$

Observe that

$$\frac{d\lambda_2^*}{d\eta} = \frac{d\lambda_3^*}{d\eta} = \frac{3\theta\phi^{4+3\eta}(1+\phi)\log\phi}{\left[\phi^{3\eta}\left(1-\phi-\phi^2\right)-\phi^3\right]^2}$$

is negative because $\log\phi < 0$. Thus, like in the 3-region case, the construction of a modern infrastructure that connects the large regions is detrimental to the in-between regions.

Let us now consider the more interesting case where *the intermediate region* 3 *is connected to the new link while region* 2 *remains unconnected.* The transportation cost matrix thus becomes:

$$\Phi = \begin{pmatrix} 1 & \phi & \phi^{2(1-\eta)} & \phi^{3(1-\eta)} \\ \phi & 1 & \phi & \phi^{2-\eta} \\ \phi^{2(1-\eta)} & \phi & 1 & \phi^{1-\eta} \\ \phi^{3(1-\eta)} & \phi^{2-\eta} & \phi^{1-\eta} & 1 \end{pmatrix}. \tag{17}$$

Since $\lambda_2^* = \lambda_3^*$ when regions 2 and 3 are unconnected, the difference between the new equilibrium firm shares $\lambda_2^{\circledast}$ and $\lambda_3^{\circledast}$ only stems from the sole existence of LHE when region 3 is connected. The new equilibrium distribution is asymmetric. Some tedious calculations show that regions 2 and 3 host different firm shares, given by the following expressions:

$$\begin{aligned} \lambda_2^{\circledast} &= \frac{2\theta-\phi}{2(1-\phi)} + \frac{2\theta\phi}{1-2\phi+\phi^{2\eta}} + \frac{\theta\phi^{1+2\eta}(\phi+\phi^\eta)}{\phi^3+\phi^{2+2\eta}-2\phi^{3+2\eta}-\phi^{3\eta}+\phi^{1+3\eta}}, \\ \lambda_3^{\circledast} &= \frac{\theta(\phi^{2+4\eta}-\phi^4-\phi^{4\eta}-\phi^{2+2\eta}+2\phi^{4+2\eta})}{(\phi^\eta-\phi)\left(\phi^3+\phi^{2+2\eta}-2\phi^{3+2\eta}-\phi^{3\eta}+\phi^{1+3\eta}\right)} - \frac{(1-2\theta)\phi^{1-\eta}}{2(1-\phi^{1-\eta})} \\ & \quad -\frac{\theta\phi}{1-2\phi+\phi^{2(1-\eta)}} - \frac{(1-2\theta)\left[\phi^{2(1-\eta)}-\phi^2\right]}{2(1-\phi)\left[1-\phi^{2(1-\eta)}\right]}. \end{aligned}$$

It is reasonable to expect that region 2 loses firms for not being connected to the new link. By contrast, it is not clear what happens to λ_3^*. Indeed, it can be shown that λ_2^* always decreases with η while λ_3^* also decreases with η only if θ and/or η is small.

Set

$$\tilde{\theta} \equiv \frac{(1-\phi)^2(1+3\phi)}{2\left(2+5\phi-\phi^2+12\phi^3\right)} \in (0, 1/4),$$

which decreases with ϕ.

The following proposition is proven in Appendix B.

Proposition 2 — *Assume an interior equilibrium. The following two cases may arise.*

(i) If $\theta < \tilde{\theta}$, *then* $\tilde{\eta} \in (0,1)$ *exists such that* $\lambda_3^* - \lambda_2^* < 0$ *holds when* $\eta \in (0, \tilde{\eta})$, *while* $\lambda_3^{\circledast} - \lambda_2^{\circledast} > 0$ *when* $\eta \in (\tilde{\eta}, 1)$.

(ii) If $\theta \geq \tilde{\theta}$, *then* $\lambda_3^{\circledast} - \lambda_2^{\circledast} \geq 0$ *holds for all* $\eta \in (0,1)$.

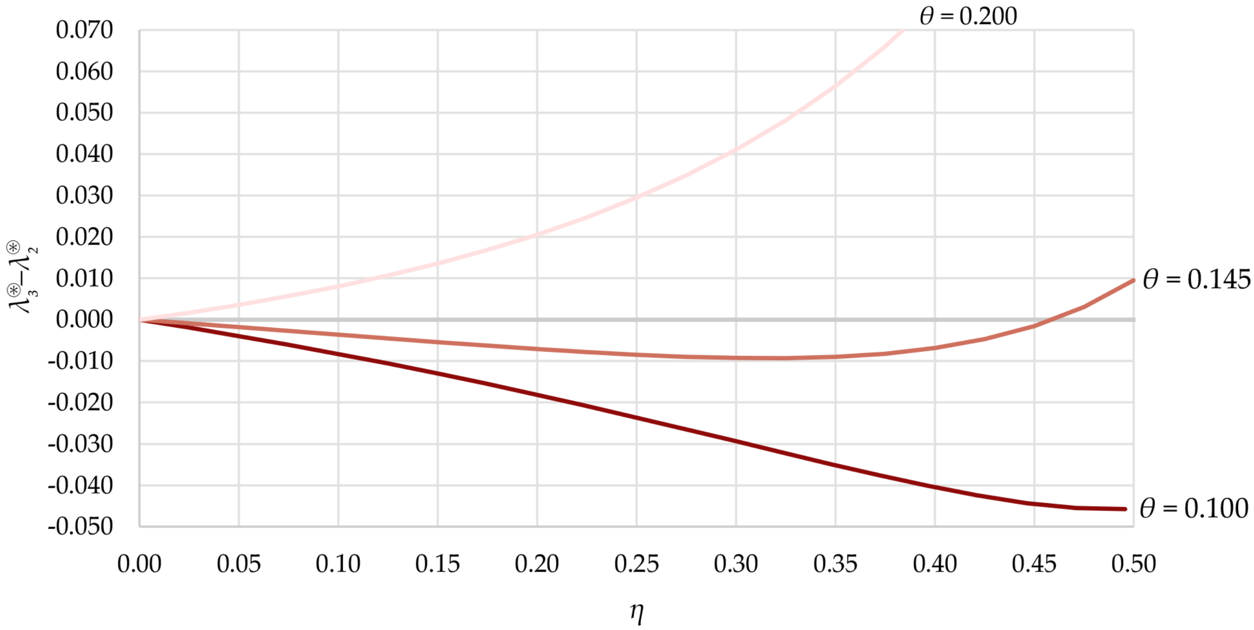

Figure 1: Long-haul economies and being connected to the network

This proposition shows that the sign of $\lambda_3^{\circledast} - \lambda_2^{\circledast}$ is a priori undetermined. Under strong LHE ($\eta > \tilde{\eta}$), the connected region always hosts a larger mass of firms than the unconnected region ($\lambda_3^{\circledast} - \lambda_2^{\circledast} > 0$), the reason being that the former has a much better access to the core regions than the latter. However, under weak LHE ($\eta < \tilde{\eta}$), being connected to the new transportation link is not sufficient for region 3 to accommodate more firms than region 2. Indeed, when the two intermediate markets are small ($\theta < \tilde{\theta}$), more firms choose to set up in the unconnected region than in the connected one because the home market effect is sufficiently strong for the core regions to pull out many firms from the connected region. By contrast, when the two intermediate regions are sufficiently large ($\theta > \tilde{\theta}$), being connected is always beneficial to region 3. All of this illustrates once more how LHE interact with the size of regions to affect the relative attractiveness of regions that are otherwise symmetric. These results are illustrated in Figure 1 for $\phi = 0.2$, implying that $\tilde{\theta} = 0.168$. This figure shows that when the intermediate region is small, for realistic values of long-haul economies, the connected region 3 will always lose firms. By contrast, if the intermediate regions are sufficiently large ($\theta > \tilde{\theta}$), being connected will always be beneficial. For intermediate cases of θ, whether region 3 will gain or lose firms depends on the size of long-haul economies η.

In the next section, we test the impact on region i of being connected to a Shinkansen line ($\eta > 0$) versus not being connected ($\eta = 0$). For this reason, we consider the log of the change in employment shares in region i given by $\log(\lambda_i^{\circledast} / \lambda_i^{\circledast}\big|_{\eta=0})$. Since $\lambda_2^{\circledast}\big|_{\eta=0} = \lambda_3^{\circledast}\big|_{\eta=0}$,the sign of

$\lambda_3^{\circledast} - \lambda_2^{\circledast}$ is the same as that of $\log(\lambda_3^{\circledast} / \lambda_3^{\circledast}\big|_{\eta=0}) - \log(\lambda_2^{\circledast} / \lambda_2^{\circledast}\big|_{\eta=0})$.Hence, Proposition 2 and the ensuing discussion remain valid when we replace $\lambda_i^{\circledast}$ with $\log(\lambda_i^{\circledast} / \lambda_i^{\circledast}\big|_{\eta=0})$.

Finally, it can be shown that $\tilde{\phi} \in (0, 1)$ exists such that $\lambda_3^{\circledast} - \lambda_2^{\circledast} > 0$ when $\phi > \tilde{\phi}$, whereas $\lambda_3^{\circledast} - \lambda_2^{\circledast} < 0$ for $\phi < \tilde{\phi}$. Therefore, a more efficient transportation system ($\phi \uparrow$) and stronger LHE ($\eta \uparrow$) due to technological progress in HSRs have similar effects on the connected and unconnected regions.

The endogeneity of market size. So far, we have chosen to isolate the effects of LHE on firms' locations from market and industry size considerations. We acknowledge that both are endogenous because consumers can migrate between municipalities. However, it is readily verified that prices in the regions connected to the new transportation infrastructure are lower than those in the unconnected ones. This could incentivize consumers to move to connected regions, thus making these regions bigger ($\theta_i \uparrow$). When the increase in market size is strong enough, Proposition 2 implies that employment in the connected regions always increases ($\lambda_i \uparrow$). Hence, when the market size is endogenous, we are inclined to find a positive effect of being connected. Nevertheless, we will show in the next section that this effect is negative. Thus, allowing for migration should not reverse our main conclusions.

5 Empirical analysis

5.1 Data

In this section, we aim to test the empirical plausibility of Proposition 2 by investigating *the effects of being connected to the Shinkansen network in Japan.* We gather data on employment and population at the municipality level between 1957 and 2014. The information on each municipality's location is provided by the *National Land Numerical Information* of the *Ministry of Land, Infrastructure, Transport, and Tourism.* Since municipal boundaries have changed several times, we have created consistent geographical units based on the definition of municipalities in 2015. Data on total employment in each municipality are obtained from the *Establishment Census* for 1957, 1972, 1978, 1981, 1986, and 1991, the *Establishment and Enterprise Census* for 1996, 2001, and 2006, and the *Economic Census for Business Frame* for 2009 and 2014. Further, we obtain the municipality-level data on population from the *Census of Population* for 1955 every five years until 2005, as well as for 2008 and 2013. We match population to the closest years for which we observe employment. We focus on municipalities on Japan's main islands, including Hokkaido, Honshu, Shikoku, and Kyushu. We end up with a sample of $1,658$ municipalities, of which three municipalities have missing data on employment in 2014.

Japan's municipalities are comparable in size to counties in the U.S. The average municipal population is about 75 thousand, while it is about 100 thousand in the U.S. Still, one may argue that municipalities cannot be considered as independent areas because workers may commute between municipalities. In Appendix C, we therefore consider bigger spatial areas, such as commuting zones (Adachi *et al.*, 2020). Furthermore, since commuting is particularly important in large metropolitan areas, we also estimate specifications in which we exclude municipalities in Tokyo, Osaka and Nagoya – the three largest metropolitan areas in Japan.

The data on the railway and expressway network is from the *National Land Numerical Information*. Expressways are considered to be high-speed controlled-access toll highways. We combine the information on expressways with information from Open Street Map in 2015 on national highways to form a complete picture of Japan's highway network. For 2014, we determine which municipalities are connected to the highway network. Since we know the opening date of each railway line, we can construct the railway networks in each year for which we have data on employment. For each municipality, we then determine for each year whether it hosts a station on the Shinkansen HSR network.[5]

For our identification strategy, to be explained later on, we also gather data on planned and historic transport networks. First, we use a plan from 1942 for a high-speed railway line to connect Tokyo to Shimonoseki and further to Beijing. This railway line was planned to transport soldiers and cargo. Second, we use the plan for the Shinkansen network that was approved by the Cabinet in around 1972. The plan was around for longer and included the Shinkansen line between Tokyo and Shimonoseki, which was opened between 1964 and 1975. We refer to the latter plan as the '1972 plan'. We further use hardcopy maps of the *National Road Plan* provided by the *Home Ministry* in 1943. The total length of highways was planned to be $5,490$ km. Further, we obtain information on roads in 1900 obtained from *National Land Numerical Information*. In Appendix C.3 we consider two alternative instruments. The first one, which we call the 'Tanaka plan', refers to the Minister of International Trade and Industry, and later the Prime Minister of Japan, Kakuei Tanaka who published a book titled *Remodelling the Japanese archipelago*. In there, Tanaka laid out a plan for the Shinkansen network. Details of the Takana plan are obtained from Sargent (1973) and various internet sources. The second alternative instrument is based on a least-cost-path spanning tree (LCPST) network proposed by Faber (2014).

Figure 2 provides a map of Japan's transportation networks in 2014. Here we observe that the Shinkansen network links essentially all large cities in Japan. We further report and discuss descriptive statistics of our data in Appendix C.2.

[5]Note that we disregard here two 'Mini'-Shinkansen lines between Fukushima and Shinjo and between Morioka and Akita. The speed on these lines (130 km/h) is the same as on the other railway lines. Therefore, they are not really HSR lines. Nevertheless, we confirm that including these two mini-Shinkansen lines leaves the results essentially unaffected.

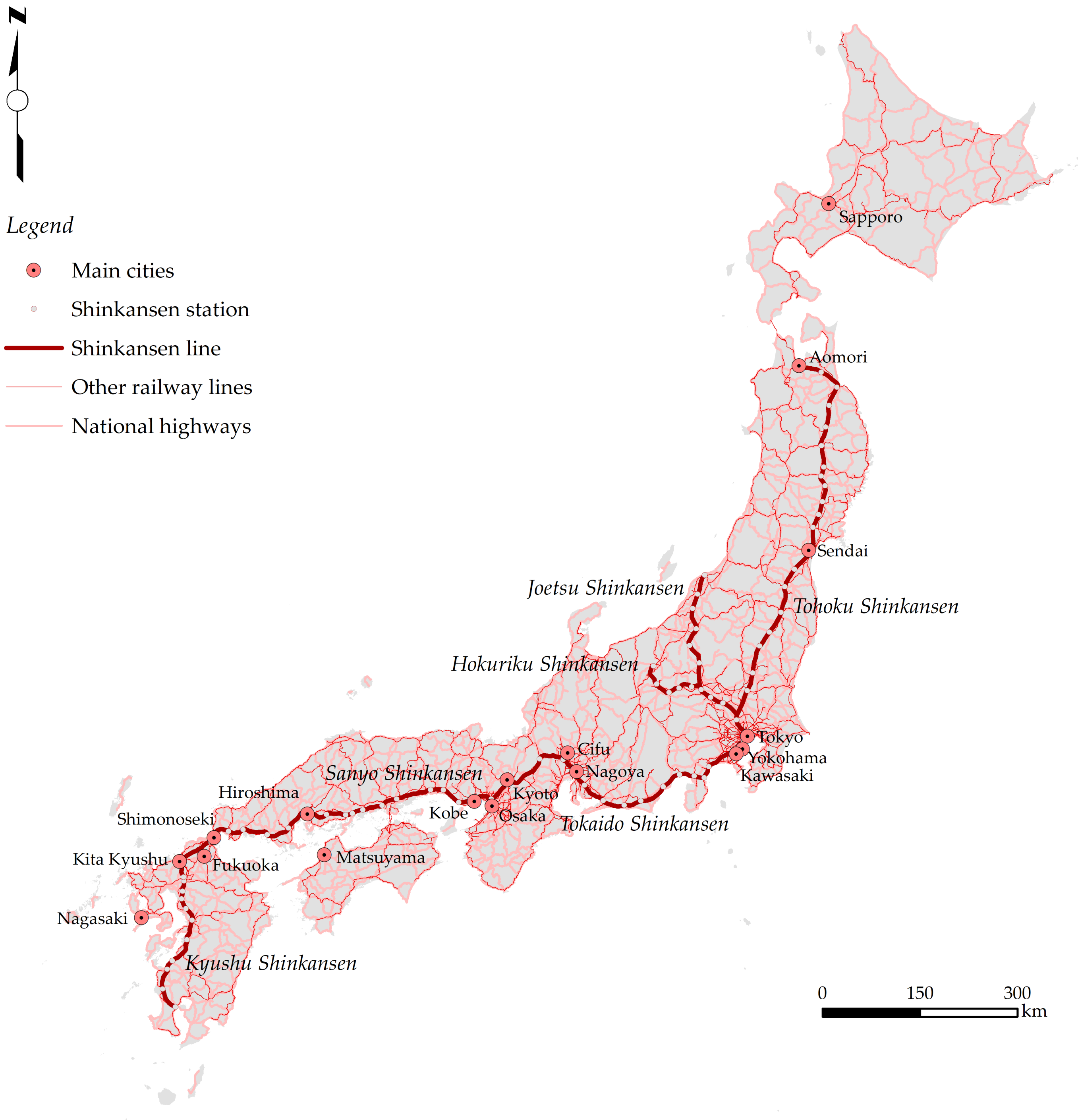

Figure 2: Overview of Japan's transportation networks

5.2 Methodology

Baseline specifications. Our baseline specification focuses on the impact of the opening of a Shinkansen station on employment in intermediate areas in order to test Proposition 2. We emphasize that we consider a 'reduced-form' approach where market access is proxied by a dummy variable indicating whether a municipality has an HSR station. Alternatively, one could choose an estimation approach using the structure of the model. We refrain from using a structural estimation approach and refer to Hayakawa *et al.* (2021) for an analysis of the effects of a major transportation infrastructure in a general equilibrium setting.

Because employment effects of infrastructure investments may take several years, or even decades, to materialize, we initially focus on long-differences specifications to increase the signal-to-noise ratio. We estimate a reduced-form equation of the log of the change in employment shares between 1957 and 2014:

$$\log \frac{\lambda_{i\bar{t}}}{\lambda_{i\underline{t}}} = \alpha + \beta S_{i\bar{t}} + \gamma h_{i\bar{t}} + \delta \log \frac{L_{i\underline{t}}}{A_i} + \zeta G_i + \varepsilon_{i\underline{t}\bar{t}}, \tag{18}$$

where $\lambda_{i\bar{t}}$ is the share of *employment* in municipality i in $\bar{t} = 2014$ and $\lambda_{i\underline{t}}$ the share of employment in $\underline{t} = 1957$; $S_{i\bar{t}}$ is a dummy that takes the value one if municipality i has a Shinkansen station at $\bar{t}$ and zero otherwise; $h_{i\bar{t}}$ is another dummy that indicates whether municipality i is connected to the highway network; $L_{i\underline{t}}/A_i$ is the population density in 1957; $G_{i\underline{t}}$ are geographic characteristics, such as climatic variables and the share of developed land; and $\varepsilon_{i\underline{t}\bar{t}}$ denotes the error term.

We are interested in a causal effect β of being connected to the Shinkansen on employment. This raises several issues. First, equilibrium employment shares depend on location-specific productivity and amenity factors, which are generally unobserved. It is very likely that the decision to build infrastructure depends on these factors. For example, particularly productive cities offering many amenities attract people and firms, which in turn attract transportation investments. Redding and Turner (2015) generally consider three approaches to mitigate endogeneity concerns: *(i)* apply an inconsequential unit approach; *(ii)* use planned route instrumental variables; and *(iii)* use historical route instrumental variables.

The first approach entails the selection of areas that are not directly targeted by the infrastructure program.[6] For example, while Toyohashi with a population of 300 thousand may have been too small to attract a dedicated Shinkansen line, the fact that it is in between Tokyo and Nagoya led Toyohashi to be connected to the Shinkansen network. In this spirit, we use the definition from 2015 of Metropolitan Employment Areas (MEAs) and Micropolitan Employment Areas (McEAs) provided by Kanemoto and Tokuoka (2002). In our final sample, there are 100 MEAs and 117

[6] This approach is essentially pursued by Ahlfeldt and Feddersen (2018), who study the impact of a HSR connection for three intermediate stops in between Frankfurt and Cologne (Germany).

McEAs. *We exclude the so-called 'central municipalities' that belong to the MEAs and McEAs because they are likely to be the reason for a Shinkansen station to have been constructed.* We therefore work with the remaining, in-between, municipalities because they are unlikely to be the reason why Shinkansen lines were planned.

The second approach, which was pioneered by Baum-Snow (2007), uses planned routes as instruments for actual infrastructure developments. The idea is that planners did not choose to connect municipalities on basis of differences in unobservable characteristics. More specifically, *we combine the inconsequential location approach, as explained above, with an instrumental variables approach based on infrastructure plans.* We consider two types of plans. The first is the 1942 plan to link Tokyo to Beijing via a fast railway line. The idea was to increase transporting passengers and goods between Japan, Korea and China. Our instrument is then a dummy indicating whether a municipality has a stop on this planned HSR. However, because this planned line only is relevant for the eastern part of Japan, our instrument may be weak. Therefore, we also consider a plan that was approved by the cabinet between 1971 and 1973, although similar versions of these plans were around for longer (Sargent, 1973).[7] We refer to this plan as the '1972 plan' and define a dummy indicating whether a municipality has a stop in the corresponding plan. Maps of the plans are reported in Appendix C.1. We do not exploit historic instruments for the Shinkansen network (the third approach in Redding and Turner, 2015), as historic instruments appear to be weak.

We consider a few generalizations of equation (18). First, one may argue that highway construction is also endogenous. We therefore estimate specifications where we instrument for $h_{i\bar{t}}$ with roads in 1900 as well as the 1943 highway plan. Second, one may be concerned that including population density in 1957 is not preferred because changes in employment density and the level of population density in 1957 may be both correlated to unobserved location characteristics. Because long-run city size is robust to large temporary shocks, we also estimate a specification where we control for the population density in the 9$^{\text{th}}$ century instead. We also control for changes in population, as to allow for migration effects. Third, we consider various spatial econometric specifications and study spatial heterogeneity, e.g., by estimating nonparametric models where the impact of a Shinkansen station depends on the population share in 1957. Fourth, we aim to exploit the full content of the data by estimating fixed effects specifications:

$$\log \lambda_{it} = \alpha_i + \alpha_{i\pi t} + \beta S_{i\bar{t}} + \gamma h_{it} + \delta \log L_{it} + \varepsilon_{it}, \tag{19}$$

[7] To be precise, the Hokuriku, Kyushu, Hokkaido lines were approved in 1973, while the Tohoku and Joetsu lines were approved in 1971. The Tokaido line between Tokyo and Shin-Osaka was already approved in 1958 and started operating in 1964, while the Sanyo line between Shin-Osaka and Hakata was approved between 1965 and 1969 and started operations from 1972 onwards.

where α_i and $\alpha_{i\pi t}$ are municipality and prefecture π-by-year fixed effects. Hence, we control for time trends at the regional level. Because we cannot control for population in 1957 (because this is time invariant), we control for population density in each year t (although the estimates are robust to excluding population density). Because it may take several years before the effects of a Shinkansen station opening materialize, the OLS fixed effects specifications have a lower signal-to-noise ratio than (18), likely leading to an underestimate of β. To address this issue, we instrument S_{it} with the 1972 plan interacted with year t. The instrument captures the fact that locations with planned stations will be more likely to receive a station during the sample period.

5.3 Results

Baseline results. We report the results from estimating equation (18) in Table 2. Column (1) is a naive regression, where we regress the log of the change in the employment share between 1957 and 2014 on whether the municipality has a Shinkansen station in 2014. Unexpectedly (at least to us), *a Shinkansen station leads to less employment in the municipality*, a finding that is in line with Proposition 2 when markets are small and LHE are not too strong, as well as with empirical results by Faber (2014). In column (2) we control for whether the municipality is connected to the highway network and the population density in 1957. If anything, the Shinkansen coefficient becomes somewhat stronger, but is not statistically significantly different. The estimate in column (2) implies that a Shinkansen station leads to a reduction in employment of $\exp(-0.246) - 1 = 21.8\%$, which is sizable. In column (3), we additionally control for geography and weather controls. This leads to a nearly identical point estimate. It also makes sense to have look at the control variables. Note that being connected to highways, which are mostly used for transporting goods rather than business travellers, also has a negative effect on employment. Although this effect has the expected sign (given the negative impact of a Shinkansen station), it is not very robust to alternative specifications. We will show later that the effect is close to zero and statistically insignificant in the fixed effects specification, as well as when we instrument for being connected to the highway network. What is more important is that the effect of having a Shinkansen station is largely insensitive to controlling for highway accessibility.[8]

In columns (4)-(6) we instrument for the Shinkansen station dummy by the 1942 and 1972 plans. The instruments appear to be strong, as the Kleibergen-Paap F-statistic exceeds 10. The first-stage results are reported in Appendix C.2. We find that a station in the 1972 plan increases the probability to have a station in 2014 by about 50%, while the 1942 plan adds an

[8] Note that population density in 1957 has a negative effect once we control for the share of land in the municipality that is developable. This means that places with a higher density in 1957 have a relatively lower employment growth. The coefficient indicates that a 100% increase in population density in 1957 lowers employment growth by about 5%.

additional 20%, but the impact of having a station in the 1942 plan is only marginally statistically significant. Going back to Table 2, we now find that the impact is considerably stronger when using instrumental variables. The coefficient in column (4) indicates that *a Shinkansen station lowers employment by* 31%. Although the impact may seem large, the order of magnitude is very much in line with empirical studies adopting similar approaches such as Faber (2014) and Baum-Snow *et al.* (2017). Also, the finding that 2SLS coefficients are considerable stronger than the OLS results is also in line with the existing literature (Baum-Snow, 2007; Faber, 2014; Garcia-López *et al.*, 2015). We attribute the (absolute) increase in the coefficient due to the particular nature of the Shinkansen network, which explicitly promoted economic growth of cities, rather than peripheral areas (Sato, 2015). In other words, it promoted cities that were likely on increasing employment trends, absent the construction of the Shinkansen. We think that the OLS estimates may not adequately control for these unobserved trends and therefore deliver underestimates of the effect of interest.

Table 2 — Regression results
(Dependent variable: the log of the change in the employment share)

				Instrumental variables		
	(1) OLS	(2) OLS	(3) OLS	(4) 2SLS	(5) 2SLS	(6) 2SLS
Shinkansen station in 2014	-0.1685*	-0.2455*	-0.2477*	**-0.3775****	**-0.3898****	**-0.3818**
	(0.0913)	(0.1293)	(0.1357)	(**0.1841**)	(**0.1858**)	(**0.5072**)
Highway in 2014		-0.1818*	-0.2081**	-0.2070**	-0.2069**	-0.3189***
		(0.1099)	(0.0943)	(0.0943)	(0.0943)	(0.1091)
Population density in 1957 *(log)*		0.2480***	-0.0736**	-0.0724**	-0.0723**	-0.0874**
		(0.0215)	(0.0334)	(0.0334)	(0.0334)	(0.0372)
Geography controls (4)	No	No	Yes	Yes	Yes	Yes
Number of observations	1,412	1,412	1,412	1,412	1,412	763
R^2	0.0005	0.0974	0.2531			
Kleibergen-Paap F-statistic				27.71	44.43	2.928

Notes: We exclude municipalities that are centers of MEAs or McEAs. Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. **Bold** indicates instrumented. In column (2) we instrument a Shinkansen station in 2014 with dummy indicating whether the municipality has a Shinkansen station planned in the 1942 plan and whether it has one planned in the 1972 plan. In column (5) we only instrument with the 1972 plan, while in column (6) we only include prefectures on East-Honshu and instrument with the 1942 plan. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Columns (5) and (6) uses either one of the instruments. In column (5) of Table 2, we show that the 1972 plan instrument is especially strong, leading to virtually the same result as in column (4). In column (6) we only use the 1942 plan to instrument for whether there is a Shinkansen station in 2014. The 1942 plan to link Tokyo to Shimonoseki and further to Beijing is only relevant for Eastern Honshu. Hence, one may argue that it does not make sense to include all of Japan. In the specification listed in column (6), we therefore only include prefectures in this area. We observe that the second-stage coefficient is very similar to the preferred specification in column (4), albeit imprecisely estimated. This is mainly because the first-stage F-statistic is just about 3. Still, we find it reassuring that the two different instruments lead to similar point estimates.

Robustness. In Table 3 we consider some robustness checks. In column (1), we include all locations, including central municipalities in the MEAs and McEAs. In line with expectations, we find less negative results because 'large' areas are expected to benefit from transport improvements (see Proposition 1).

Table 3 — Robustness
(Dependent variable: the log of the change in the employment share)

	(1) 2SLS	(2) 2SLS	(3) 2SLS	(4) 2SLS	(5) 2SLS	(6) 2SLS	(7) 2SLS
	All locations	Weighted regression	+ Railway connections	Base year 1972	+ Exclude realized	Instrument for highways	Population in 900
Shinkansen station in 2014	**-0.1643***	**-0.3775****	**-0.3952****	**-0.3836****	**-0.2147**	**-0.4313****	**-0.3769****
	(**0.0847**)	(**0.1841**)	(**0.1865**)	(**0.1677**)	(**0.1818**)	(**0.1983**)	(**0.1729**)
Δ Railway station			0.1335**				
			(0.0551)				
Highway in 2014	-0.2180**	-0.2070**	-0.2169**	-0.1228**	-0.1657**	0.7439	-0.2212**
	(0.0920)	(0.0943)	(0.0941)	(0.0608)	(0.0809)	(0.5282)	(0.0941)
Population density in 1957 *(log)*	-0.0679**	-0.0724**	-0.0695**	-0.0690***	-0.1063***	-0.0540	
	(0.0279)	(0.0334)	(0.0334)	(0.0233)	(0.0332)	(0.0351)	
Population density in 900 *(log)*							0.1465***
							(0.0221)
Geography controls (4)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	1,655	1,412	1,412	1,412	804	1,412	1,408
Kleibergen-Paap F-statistic	223.8	27.71	27.69	14.97	20.47	8.115	27.94

Notes: We exclude municipalities that are centers of MEAs or McEAs. Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. **Bold** indicates instrumented. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In column (2) of Table 3, we aim to address the concern that our results are driven by a few small municipalities. Because we take the log of employment growth, implying that we look at relative changes, we might put too much emphasis on small municipalities where relative changes in employment can be large. We then estimate weighted regressions, where the weights depend on the population in a municipality in 1957. We find very similar effects.

In column (3), we additionally control for changes in whether a municipality is connected to the ordinary railway network. Overall, we see a small increase of 4.3% in the share of municipalities that have a railway connection. However, some railway lines have been closed. For example, on Hokkaido we witness a decrease of 27% between 1957 and 2014 in the municipalities that are connected to the railway network. The coefficient in column (2) indicates that ordinary railway connections generally attract employment. The connection to an ordinary railway line is associated with an increase in employment of about 14%. However, as we do not instrument for railway connections, we should be careful to interpret this effect as a causal effect because, especially in declining municipalities, railway lines have been closed. Furthermore, railway lines may pick up decentralization effects due to lower commuting costs in suburban areas (Baum-Snow *et al.*, 2017). By contrast, Shinkansen trains are hardly used for commuting. What is more important is that the negative effect of receiving a Shinkansen station hardly changes.

In column (4) of Table 3, we choose as base year 1972 instead of 1957. As most of the Shinkansen came in operation mostly after 1972, we expect the effects to have been materialized after 1972. Hence, the coefficient are not expected to change much. Indeed, the estimated coefficient (-0.384) is very similar to the preferred specification listed in column (4) of Table 2. In column (5) we improve on our 1972 plan instrument by excluding municipalities that are crossed by the Tokaido-Sanyo line (which was already partly in operation before 1972). At the same time, we focus on employment changes between 1972 and 2014. In this way, we ensure that the plan was really preceding any changes in employment, rather than potentially incorporating these. The coefficient is somewhat lower (partly because some of the most affected municipalities are on the Tokaido-Sanyo line). However, the standard error is quite large, implying that the effect is not statistically significantly different from the preferred specification.

One may further argue that the highway variable is endogenous. Hence, in column (6) we instrument for highways with the highway plan from 1943 and roads from 1900. The Kleibergen-Paap F-statistic indicates that the instruments are sufficiently strong. We find that the negative impact of the Shinkansen on employment is now even somewhat stronger, as the coefficient indicates that a Shinkansen station lowers employment by 35%. The impact of highways is now positive, but very imprecisely estimated. Hence, we cannot draw firm conclusions about the effect of highways.

Column (7) investigates whether controlling for population in 1957 matters. Using historic data from Kito (1996) and Takada (2017), we instead control for the population in the 9$^{\text{th}}$ century.[9] We find a very comparable effect of a Shinkansen station.[10]

In Appendix C.3 we consider the use of alternative instruments, i.e., the Tanaka plan, and we construct a least-costs-path spanning tree network as an instrument. We show that the results are

[9] More specifically, the data on the local population in 900 are taken from Kito (1996). Although the estimates of population are available for 68 provinces in 900, the provincial boundaries are obviously different from current municipal boundaries. We address this issue by distributing the population in each province according to the share of land of each municipality in the corresponding historical province. We note that Kito (1996) does not provide information on the population in Hokkaido. By using the information on the number of archaeological remains, Takada (2017) estimated the population size in Hokkaido around the 9$^{\text{th}}$ century as 37 thousand. This number is distributed to each municipality in Hokkaido based on its land share.

[10] We also control for the change in population density between 1957 and 2014. This does not affect the results, which are available upon request.

somewhat stronger, albeit not significantly so. However, the instruments are considerably weaker than the 1942 and 1972 plans, leading to imprecise estimates, so we prefer instruments based on these latter plans. In Appendix C.4, we use commuting zones instead of municipalities and show that the main results are robust. Appendix C.5 tests for robustness of the results by *(i)* including McEA centers, *(ii)* excluding the three largest MEAs (Tokyo, Osaka and Nagoya), *(iii)* excluding municipalities close to areas that have received a Shinkansen station, *(iv)* controlling for changes in population to address the issue that population may not be given, but is endogenous, and *(v)* exclude outliers in the change in employment. The results are robust.

In Appendix C.6 we re-examine our OLS results by investigating whether omitted variable bias is important. Using Oster's (2019) GMM approach, we show that, conditional on our set of controls, omitted variable bias is not a big issue.

Spatial heterogeneity. We further investigate whether spatial heterogeneity is important.

As a first step we estimate spatial econometric models. As is common practice, we construct a row-standardized inverse distance weight matrix. The most obvious candidate for a model like ours is probably the spatial cross-regressive model that allows for the effects of Shinkansen station in nearby municipalities (as defined by the weight matrix). In column (1), we instrument only for having a Shinkansen station using the 1942 and 1972 plans. The spatial lag is negative and statistically significant. For a standard deviation increase in the spatial lag of the Shinkansen station, employment decreases by 5%. Hence, also stations in the direct vicinity seem to generate a negative employment effect, which makes sense. Interestingly, the main effect of interest is hardly affected.

One may argue that the spatial lag of the Shinkansen station dummy might be endogenous as well. In column (2), we therefore instrument the spatial lag of the Shinkansen station dummy with the spatial lag of the 1972 and 1942 plans. These instruments seem sufficiently strong and the spatial lag effect is again negative and statistically significant. The magnitude of the coefficient is very similar to the previous specification. Furthermore, the effect of interest is only marginally statistically significant, although, at the same time, it is not statistically significantly different from the preferred baseline estimate. We think it is safe to conclude that these results indicate that the negative effect we measure does not capture the effect of employment moving just across the municipality boundary – otherwise the spatial lag of the Shinkansen station dummy would indicate a positive effect.

In column (2) of Table 4, we allow residuals to be spatially correlated. The spatial error effect is highly statistically significant, confirming that residuals are indeed spatially correlated. We find that the effect becomes slightly lower and turns statistically insignificant. However, this is an issue of precision, as the coefficient is sizably negative and close to the baseline specifications.

For completeness, column (3) includes a spatial lag of the dependent variable. We are not great proponents of spatial lag models because the spatial lag just may just be seen as a reduced-form spatial cross-regressive model, while the marginal effects of a Shinkansen station may be harder to interpret (Gibbons and Overman, 2012). However, we still think that the spatial lag model might be useful in showing the robustness of our results. It is reassuring that the coefficient of interest hardly changes, although the spatial lag is strong and statistically significant.

Table 4 — Spatial econometric models
(Dependent variable: the log of the change in the employment share)

	(1)	(2)	(3)	(4)	(5)
	2SLS	2SLS	GMM	GMM	GMM
	Spatial cross-regressive model	Spatial cross-regressive model	Spatial error model	Spatial lag model	All spatial effects
Shinkansen station in 2014	**-0.3677****	**-0.3106***	**-0.2745**	**-0.2700**	**-0.2481**
	(0.1861)	**(0.1864)**	**(0.1852)**	**(0.1919)**	**(0.1906)**
Highway in 2014	-0.2033**	-0.2033**	-0.2009**	-0.1989**	-0.1990**
	(0.0943)	(0.0943)	(0.0912)	(0.0883)	(0.0879)
Population density in 1957 *(log)*	-0.0701**	-0.0703**	-0.0917***	-0.0968***	-0.0992***
	(0.0333)	(0.0333)	(0.0346)	(0.0339)	(0.0342)
Spatial effects:					
W· Shinkansen station in 2014	-8.2280***	**-9.2483*****			**1.1664**
	(2.4896)	**(3.2142)**			**(3.5448)**
$\mathbf{W} \cdot \epsilon$			1.4377***		-0.0040
			(0.3320)		(0.4306)
$\mathbf{W} \cdot \log \frac{\lambda_{2014}}{\lambda_{1957}}$				1.2399***	1.3371***
				(0.1749)	(0.2002)
Geography controls (4)	Yes	Yes	Yes	Yes	Yes
Number of observations	1,412	1,412	1,412	1,412	1,412
Kleibergen-Paap F-statistic	27.79	14.78			

Notes: **W** is a row-standardized inverse distance-weight matrix. We exclude municipalities that are centers of MEAs or McEAs. Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. **Bold** indicates instrumented. We instrument a Shinkansen station in 2014 with dummy indicating whether the municipality has a Shinkansen station planned in the 1942 plan and whether it has one planned in the 1972 plan. In columns (2) and (5) we also instrument for **W**·Shinkansen station in 2014 with the spatial lag of the 1972 plan and the spatial lag of the 1942 plan. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Column (4) in Table 2 includes all spatial effects, showing that the main effect is very similar to the preferred specification in column (4) of Table 2 but just not significant at conventional levels. We now find a small positive effect related to the spatial lag of the Shinkansen station dummy, capturing the opening of a Shinkansen station in nearby municipalities, although it is very imprecise, which can be explained by correlation between the spatial lag of changes in employment and the spatial lag of the Shinkansen station dummy.

Second, in Appendix C.7 we estimate non-parametric regressions where we let the impact of being connected be dependent on the population share in 1957. We find that the effect is the most negative for small municipalities, while it is monotonically increasing in population share $\theta_{i\underline{t}}$, which is line with predictions from Proposition 2. However, the confidence bands are somewhat wide, which precludes us from drawing firm conclusions.

Third, so far we are agnostic about where employment from connected places moves to. It may well be that firms move within metropolitan areas to more central areas where the home market effect is stronger. One may wonder whether we can say something on what type of areas have benefited from the opening of Shinkansen stations. Our model does not make strong predictions on this, and we do not pursue an approach in this paper where we estimate spatial general equilibrium models that are much more suitable to answer these type of questions. Still, we provide suggestive evidence in Appendix C.8 that particularly suburban and central municipalities seem to have benefited from the opening of Shinkansen stations. We find that those areas see an employment growth of about 30% compared to municipalities that have received a station. By contrast, remote areas and municipalities in McEAs lose employment relative to municipalities that have been connected to the Shinkansen network.

Fixed effects estimation. One may argue that by only including 1957 and 2014 in the analysis we disregard most of the data.[11] We argued that the signal-to-noise ratio is lower when including intermediate years because it may take years, or even decades, before spatial effects of infrastructure changes materialize. Hence, fixed effects estimations including intermediate years may lead to a bias in the estimates towards zero. Still, one might argue that if results would be very different, this would be suspicious. We therefore essentially replicate the baseline results in Table 2, but instead estimate equation (19). We note that we do not have time-varying data on the opening of national highways; hence we focus just on expressways.

Column (1) in Table 5 is a somewhat naive specification where we only control for municipality and year fixed effects. We confirm the negative effect that we find earlier, although as expected, the impact is somewhat lower. In column (2) we include controls, i.e., whether a municipality has an expressway connection and the log of population density. Furthermore, because we have a much larger number of observations, we can also control in a better way for region-specific trends by including prefecture-by-year fixed effects, which implies that we identify the effects of a Shinkansen station within prefectures. The coefficient implies that a Shinkansen station lowers employment by 12.8%. Note that the effect of expressways is essentially zero with a rather small standard error. Hence, in this specification, expressways do not seem to lead to a lower employment share. Population density has an elasticity that is essentially equal to 1, which is in line with the idea that commuting is not really an issue.[12] In column (3), we further improve by

[11] Recall that we have data on employment from 1957, 1972, 1978, 1981, 1986, 1991, 1996, 2001, 2006, 2009, and 2014.

[12] To illustrate this, consider a situation where people would live at very different locations than where they work. We would then find an elasticity that would be considerably lower than 1, because an increase in population would

including linear trends of the geography controls, leading to a nearly identical coefficient.

Table 5 — Regression results – fixed effects estimation
(Dependent variable: the log of the employment share)

				Instrumental variables	
	(1)	(2)	(3)	(4)	(5)
	OLS	OLS	OLS	2SLS	2SLS
Shinkansen station	-0.0930	-0.1372***	-0.1275***	**-0.5080*****	**-0.2837*****
	(0.0610)	(0.0381)	(0.0380)	(**0.1259**)	(**0.1032**)
Expressway connection		-0.0083	-0.0059	-0.0043	-0.0224*
		(0.0109)	(0.0110)	(0.0111)	(0.0134)
Population *(log)*		0.9993***	0.9915***	0.9902***	1.0150***
		(0.0207)	(0.0238)	(0.0240)	(0.0336)
Geography controls×year trends	No	No	Yes	Yes	Yes
Prefecture×year fixed effects	No	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Municipality fixed effects	Yes	Yes	Yes	Yes	Yes
Number of observations	15,562	15,562	15,562	15,562	8,067
R^2	0.9467	0.9801	0.9803		
Kleibergen-Paap F-statistic				42.68	22.28

Notes: We exclude municipalities that are centers of MEAs or McEAs. Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. **Bold** indicates instrumented. In column (4) we instrument a Shinkansen station in 2014 with dummy indicating whether the municipality has a Shinkansen station planned in 1972 interacted with the year of observation. In column (5) we exclude municipalities on the Tokkaido-Sanyo line. Standard errors are clustered at the municipality level and in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In columns (4) and (5), we instrument for the Shinkansen station dummy. Because the 1972 plan is time invariant, we interact the dummy indicating whether a municipality has a station in the 1972 plan with the year of observation. In other words, we expect that over time the probability to receive a station is positively correlated to whether the location has a planned station. An additional benefit of instrumenting for the Shinkansen station dummy is that the instrument is likely to be uncorrelated to the measurement error caused by uncertainty about the timing of the effect. We disregard the 1942 plan here because this appears to be a very weak instrument. Column (4) shows that the estimated effect is similar to the baseline result based on long-differences, which is reassuring. One may be concerned that the 1972 plan is preceded by observations in 1957. Furthermore, part of the Tokaido-Sanyo line has been in operation since 1964. In column (5) we therefore exclude municipalities that are on the Tokaido-Sanyo line and only include data from 1972 onwards. In line with the comparable specification in column (4) of Table 3, we find a somewhat lower effect: a Shinkansen station decreases local employment by 25%, which is very similar to the baseline estimate.

In sum, the results of different specifications deliver a consistent picture: *in-between areas lose employment when they are connected to the Shinkansen network.* The effects are sizable, ranging from about 10 to 40%.

not necessarily lead to an increase in employment.

6 Conclusion

In this paper, we have shown that the connection to a new transportation infrastructure need not foster the creation of jobs in intermediate and remote regions. In particular, we have seen that the size of local markets – which can be viewed as a proxy for regional characteristics that complement investments – is critical in determining the impact of transportation infrastructure. Furthermore, by highlighting the role of LHE associated with modern transportation modes, this paper contributes to a better understanding of why the high payoffs expected from the provision of such infrastructure do not necessarily materialize. Finally, we have used Japanese data on employment, population and the Shinkansen network to show that connecting intermediate or remote areas to the main network is unlikely to trigger the economic take-off of these areas. More specifically, we have found that, relative to areas that remain unconnected, the Shinkansen may have caused decreases in employment growth. This seems to contradict Ahlfeldt and Feddersen (2018). However, one should keep in mind that the places connected to the Cologne–Frankfurt HSR are in between two of the five top German cities, and thus may benefit from this locational advantage. In line with Hayakawa *et al.* (2021), our results highlight that the relative position of municipalities within transport networks and their underlying location fundamentals are essential in understanding why the effects of an extensive infrastructure are positive or negative. We live in a spiky world for reasons that have little to do with an insufficient provision of transportation infrastructure, at least in developed countries (Brakman *et al.*, 2020). Consequently, one should question the widespread approach that favors 'hard' investments, such as transportation infrastructures, over 'soft' ones (human capital and research capacity) in regional policy.

References

[1] Adachi, D., Fukai, T., Kawaguchi, D., and Saito, Y. (2020) Commuting zones in Japan. RIETI Discussion Paper 20-E-021.

[2] Aggarwal, G. (2018) Do rural roads create pathways out of poverty? Evidence from India. *Journal of Development Economics* 133: 375-95.

[3] Ahlfeldt, G.M. and Feddersen, A. (2018) From periphery to core: Measuring agglomeration effects using high-speed rail. *Journal of Economic Geography* 18: 355-90.

[4] Asher, S. and Novosad, P. (2020) Rural roads and local economic development. *American Economic Review* 110: 797-823.

[5] Baldwin, R.E., Forslid, R., Martin, P., Ottaviano, G.I.P., and Robert-Nicoud, F. (2003) *Economic Geography and Public Policy.* Princeton, NJ: Princeton University Press.

[6] Baum-Snow, N. (2007) Did highways cause suburbanization? *Quarterly Journal of Economics* 122: 775-805.

[7] Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M.A., and Zhang, Q. (2017) Roads, railroads, and decentralization of Chinese cities. *Review of Economics and Statistics* 99: 435-48.

[8] Behrens, K., Lamorgese, A.R., Ottaviano, G.I.P., and Tabuchi, T. (2007) Changes in transport and non-transport costs: Local vs global impacts in a spatial network. *Regional Science and Urban Economics* 37: 625-48.

[9] Blanquart, C. and Koning, M. (2017) The local economic impacts of high-speed railways: Theories and facts. *European Transport Research Review* 9: 1-14.

[10] Bogart, D. (2014) The transport revolution in industrializing Britain: A survey. In: *The Cambridge Economic History of Britain 1700 to 1870*, edited by R. Floud and J. Humphries, 368-391. Cambridge University Press.

[11] Boyer, K.D. (1997) *Principles of Transportation Economics.* Reading: Addison-Wesley.

[12] Brakman, S., Garretsen, H., and van Marrewijk, C. (2020) *An Introduction to Geographical and Urban Economics. A Spiky World.* Third Edition. Cambridge University Press.

[13] Brook, T. (2008) *Vermeer's Hat: The Seventeenth Century and the Dawn of the Global World.* Bloomsbury.

[14] Campate, F.R. and Yanagizawa-Drott, D. (2018) Long-range growth: Economic development in the global network of air links. *Quarterly Journal of Economics* 133: 1395-458.

[15] Corrado, L. and Fingleton, B. (2012) Where is the economics in spatial econometrics? *Journal of Regional Science* 52: 210-39.

[16] Couture, V., Duranton, G., and Turner, M.A. (2018) Speed. *Review of Economics and Statistics* 100: 725-739.

[17] De Best, R. (2020)ww Infrastructure spending as share of GDP in 48 OECD countries 2018. *Statista.com.*

[18] Donaldson, D. and Hornbeck, R. (2016) Railroads and American economic growth: A "market access" approach. *Quarterly Journal of Economics* 131: 799-858.

[19] Donaldson, D. (2018) Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review* 108: 899-934.

[20] Ethier, W. (1982) National and international returns to scale in the modern theory of international trade. *American Economic Review* 72: 389-405.

[21] Faber, B. (2014) Trade integration, market size, and industrialization: Evidence from China's National Trunk Highway System. Review of Economic Studies 81: 1046-70.

[22] Duranton, G. and Turner, M.A. (2012) Urban growth and transportation. *Review of Economic Studies* 79: 1407-40.

[23] Garcia-López, M.À., Holl, A., and Viladecans-Marsal, E. (2015) Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities. Journal of Urban Economics 85: 52-67.

[24] Gibbons, S., Lyytikäinen, T., Overman, H.G., and Sanchis-Guarner, R. (2019) New road infrastructure: The effects on firms. *Journal of Urban Economics* 110: 35-50.

[25] Gibbons, S. and Overman, H.G. (2012) Mostly pointless spatial econometrics? *Journal of Regional Science* 52: 172-91.

[26] Hayakawa, K., Koster, H.R.A., Tabuchi, T., and Thisse, J.-F. (2021) High-speed rail and the spatial distribution of economic activity: Evidence from Japan's Shinkansen. RIETI Discussion Paper 21-E-003.

[27] Kanemoto, Y. and Tokuoka, K. (2002) Proposal for the standards of metropolitan areas of Japan (in Japanese). *Oyo Chiikigaku Kenkyu* 7, 1-5.

[28] Kito, H. (1996) The regional population of Japan before the Meiji Period (in Japanese). *Jochi Keizai Ronsyu* 41: 65-79.

[29] Krugman, P. (1980) Scale economies, product differentiation, and the pattern of trade. *American Economic Review* 70: 950-59.

[30] Krugman, P. (1993) The hub effect: Or, threeness in international trade. In *Theory, Policy and Dynamics in International Trade*, edited by W. J. Ethier, E. Helpman, and J. P. Neary, Cambridge, MA: Cambridge University Press.

[31] McCann, P. (2005) Transport costs and new economic geography. *Journal of Economic Geography* 5: 305-18.

[32] Monte, F., Redding, S.J., and Rossi-Hansberg, E. (2018) Commuting, migration, and local employment elasticities. *American Economic Review* 108: 3855-90.

[33] Mori, T. (2012) Increasing returns in transportation and the formation of hubs. *Journal of Economic Geography* 12: 877-97.

[34] Oster, E. (2019) Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business Economics and Statistics* 37(2), 187-204.

[35] Proost, S., Dunkerley, F., van der Loo, S. Adler, N., Bröcker, J., and Korzhenevych, A. (2014) Do the selected Trans European transport investments pass the cost benefit test? *Transportation* 41: 107-32.

[36] Qin, Y. (2017) 'No county left behind?' The distributional impact of high-speed rail upgrades in China. *Journal of Economic Geography* 17: 489-520.

[37] Quinet, E. and Vickerman, R. (2005) *Principles of Transport Economics.* Cheltenham: Edward Elgar.

[38] Redding, S.J. and Turner, M.A. (2015) Transportation costs and the spatial organization of economic activity. In: *Handbook of Regional and Urban Economics. Volume 5*, edited by G. Duranton, J.V. Henderson, and W.C. Strange, 247-348. Amsterdam: Elsevier.

[39] Samuelson, P.A. (1954) The transfer problem and transport cost, II: Analysis of effects of trade impediments. *Economic Journal* 64: 264-89.

[40] Sargent, J. (1973) Remodelling the Japanese archipelago: The Tanaka plan. *Geographical Journal* 139: 426-35.

[41] Sato, N. (2015) *History of Shinkansen* (In Japanese). Tokyo: Chuokoron-Shinsha.

[42] Subrahmanyam, S. (2012) *The Portuguese Empire in Asia, 1500-1700: A Political and Economic History.* Wiley-Blackwell, 2012.

[43] Tchetgen Tchetgen, E.J. (2014) A note on the control function approach with an instrumental variable and a binary outcome. *Epidemiologic Methods* 3: 107-12.

Appendix A. Long-haul economies: Data and descriptive statistics

The data on Japan's railway network is from the *National Land Numerical Information*. For each railway line we know the opening and closing date, so that we can construct the railway network in 2014. From the *JTB Timetable* and the *JR Timetable* (*Kotsu Shinbunsha*), we obtain information on the average speeds on railways. On regular railway lines, the average speed is 60 km/h in 2014.[13] As for the Shinkansen, its speed varies from 130 km/h to 250 km/h. We then compute the fastest travel path between any two locations. We also have data on the road network from the *National Land Numerical Information* and on speeds for various types of roads from the *Road Traffic Census*. Each of Japan's $1,754$ municipalities are then assigned to either the railway or road network. Next, for each municipality pair we calculate the the bilateral *travel time* and *network distance* by rail and road.

For the Netherlands, we obtain information on the street network in 2012 from *Spinlab*, which provides information on average free-flow speeds per short road segment (the median length of a segment is 96 m), which are usually well below the speed limit due to traffic lights, roundabouts and intersections. We calculate travel times and network distances by assigning $4,033$ postcodes to the road network and then calculate the corresponding time and network distance for each pair. We further obtain travel time and network distances between all stations in the Netherlands in 2010 from the *Dutch Railways*. We calculate the travel time as the sum of the in vehicle time and the transfer time multiplied by the number of transfers.

We report descriptive statistics of the data on travel times and network distances in Table A1. The average travel time by road in Japan is about 14 hours, while the distance between municipality pairs is 724 km. Maybe surprisingly, the travel time by rail is considerably shorter on average due to the presence of high-speed rail, while the average distance between pairs is slightly higher than for the road network. The correlation between road travel time and road network distance is 0.951, while the correlation between road travel time and railway travel time is 0.936.

For the Netherlands, the average travel times are much lower because the Netherlands is a much smaller country. For example, the average travel time by road between postcode pairs is just over 1.5 hours. The correlation between road network distance and travel time is 0.893. In the Netherlands, the average travel time by rail is considerably higher than the travel time by road. This is due to the lack of high-speed trains.

[13] These speeds are computed by dividing the route-distance by the actual time when leaving Tokyo station for Shimonoseki station by local trains.

Table A.1 — Descriptive statistics

Panel A: Japan	(1) mean	(2) sd	(3) min	(4) max
Total travel time by road *(min)*	889.0294	699.0419	0.0000	5,758.8047
Total distance by road *(km)*	724.9446	494.0695	0.0000	3,031.8757
Total travel time by rail *(min)*	515.9478	450.1386	0.0000	3,731.0383
Total distance by rail *(km)*	836.2520	559.3324	0.0000	3,419.0186
Panel B: The Netherlands	**(1) mean**	**(2) sd**	**(3) min**	**(4) max**
Total travel time by road *(min)*	97.9796	54.2438	0.0000	632.3378
Total distance by road *(km)*	133.3259	69.0109	0.0000	412.2625
Total travel time by rail *(min)*	168.2285	89.4154	1.5000	743.0000
Total distance by rail *(km)*	156.3438	84.5337	0.0000	477.1000

Notes: For road and travel times in Japan we have information on 3,076,388 municipality pairs. For the road network in the Netherlands we have 16,208,683 postcode pairs, while for the Dutch railway network we have 144,780 station pairs.

Appendix B. Proof of Proposition 2

When η is slightly below $\tilde{\eta}$, we have $\lambda_2^* > \lambda_3^*$. Set $f \equiv \phi^\eta$ and let $\phi = g(f)$ be the solution to $\lambda_2^*(\phi, f) = \lambda_3^*(\phi, f)$. This function can be shown to be single-peaked or decreasing over the interval of $[\phi, 1)$. The equation $\phi = g(\phi)$ has a unique solution $\tilde{\phi}$, which is the intersection point of $\phi = g(f)$ and $\phi = f$ in the (ϕ, f) plane. Then, we differentiate $g(f)$ six times and check the corresponding derivatives at $f = \phi$ and $f = 1$. Depending on their signs, we obtain three thresholds of θ that depend on ϕ. Comparing these thresholds yields four cases: three of them lead to a single-peaked function $g(f)$ and the last one to a monotone decreasing function $g(f)$. This implies that either (i) $\tilde{f} \in (\phi, 1)$ exists such that $\lambda_2^* > \lambda_3^*$ for $f \in (\phi, \tilde{f})$ and $\lambda_2^* < \lambda_3^*$ for $f \in (\tilde{f}, 1)$ or (ii) $\lambda_2^* < \lambda_3^*$ for all $f \in (\phi, 1)$. Q.E.D.

Appendix C. Other empirical results

C.1 Descriptive statistics and maps

In Table C.1 we report descriptives for the main variables in the long-differences specifications. The largest share of employment is about 6.5% for the municipality of Osaka. We do not find a high employment share for Tokyo because Tokyo is divided in multiple municipalities. The largest absolute increase in the share of employment can be found in Yokohama (i.e., 0.8 percentage points), while the largest decrease we observe in Osaka (i.e. -2.7 percentage point). Observe that 5% of the municipalities received a Shinkansen station between 1957 and 2014, while 10%

of the municipalities now have a Shinkansen line crossing their municipality. By contrast, 93% of the municipalities are connected to the highway network. The share of developable land (which we obtain from Hayakawa *et al.*, 2021) is considerably higher than other estimates of 'inhabitable land', which would be 33% according to the *Social Indicators* by Prefecture from the *Statistics Bureau*. Inhabitable land excludes forests, lakes and other water bodies and does not use information on slopes. However, forests are technically developable, although they are often protected. Hence, we think the share of developable land is best interpreted as an upper bound estimate of the amount of actually developable land in Japan.

TABLE C.1 — DESCRIPTIVE STATISTICS IN CHANGES

	(1) mean	(2) sd	(3) min	(4) max
Change in share of employment *(in %)*	0.0000	0.0948	-2.671	0.826
Shinkansen station in 2014	0.0495	0.217	0	1
Shinkansen line in 2014	0.103	0.303	0	1
Highway in 2014	0.928	0.259	0	1
Population density in 1957 *(in km^2)*	662.5	2,041	4.280	31,840
Shinkansen station in 1972 plan	0.0627	0.243	0	1
Shinkansen station in 1942 plan	0.0115	0.106	0	1
Share of developable land in municipality	0.843	0.154	0.191	1.000
Precipitation *(in mm per km^2)*	2,345	480.6	110.2	4,065
Mean temperature in January *($^o C$)*	-5.091	5.849	-25.90	5.300
Mean temperature in July *($^o C$)*	29.03	2.633	18.40	32.80

Notes: The number of observations is 1,655.

In Table C.2 we report the full panel dataset. Because stations were built during the studied period, the share of observations with a Shinkansen station is lower than in the long-differences sample (0.031 vs. 0.050 respectively).

TABLE C.2 — DESCRIPTIVE STATISTICS FOR ALL YEARS

	(1) mean	(2) sd	(3) min	(4) max
Share of employment in total employment *(in %)*	0.0603	0.195	0	6.591
Shinkansen station	0.0312	0.174	0	1
Shinkansen line	0.0728	0.260	0	1
Railway connection	0.781	0.414	0	1
Highway in 2014	0.940	0.238	0	1
Population density *(in km^2)*	1,012	2,375	2.339	31,840

Notes: The number of observations is 18,238.

We report maps of the planned Shinkansen network in Figure 3. The idea behind the 1942 plan was to link Tokyo to Shimonoseki and even further to Beijing. The 1972 plan that was approved by the Cabinet incorporates many of the lines that are currently in operation or under construction. The plan did neither include the Chuo MagLev line between Tokyo and Nagoya, which is now under construction, nor dit it incorporate the 'Mini'-Shinkansen lines between Fukushima and Shinjo and between Morioka and Akita, which were approved in 1987 and 1990, respectively (but are not included in our analysis).

C.2 First-stage results

We report first-stage results in Table C.3. In columns (1)-(3) we focus on areas that had a Shinkansen station in 2014. We find that having a station in the 1972 plan increases the probability to have a station in 2015 by about 50 percentage points. If we only use the 1972 plan instrument in column (2), we find a similar impact. In column (3) we only include Eastern Honshu when focusing on the 1942 plan instrument. This substantially reduces the R^2 and, while the impact of the Shinkansen station in the 1942 plan is sizable, it is only marginally statistically significant.

TABLE C.3 — FIRST-STAGE RESULTS
(Dependent variable: Shinkansen station or line in 2014)

	(1) OLS	(2) OLS	(3) OLS
Shinkansen station in 1972 plan	0.5129***	0.5219***	
	(0.0783)	(0.0783)	
Shinkansen station in 1942	0.1928*		0.3774*
	(0.1093)		(0.2206)
Geography controls (4)	Yes	Yes	Yes
Number of observations	1,415	1,415	763
R^2	0.4642	0.4565	0.0578

Notes: We exclude municipalities that are centers of MEAs or McEAs. We exclude municipalities with a station in 2014 in columns (4)-(6). In columns (3) and (6) we only include prefectures on East-Honshu and instrument with the 1942 plan. Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

C.3 Other instruments

We consider here two alternative instruments for whether a municipality is connected to the Shinkansen network in 2014. According to the Tanaka plan, the Shinkansen network would have been extended to a total length of 9000 km. The extension of the Shinkansen network would be accompanied by a substantive modernization of the Japanese National Railways, including electrification and upgrading to dual gauge railways. In Figure 4 we show the lines listed in the Tanaka plan. It shows that many have been built, although there are a few lines that have not been yet considered (e.g., the line linking Shikoku with Kyushu). An issue with using the Tanaka plan as an instrument is that the locations of Shinkansen stations were not indicated, except for the main cities (which are part of centers of MEAs and, therefore, not included in our main

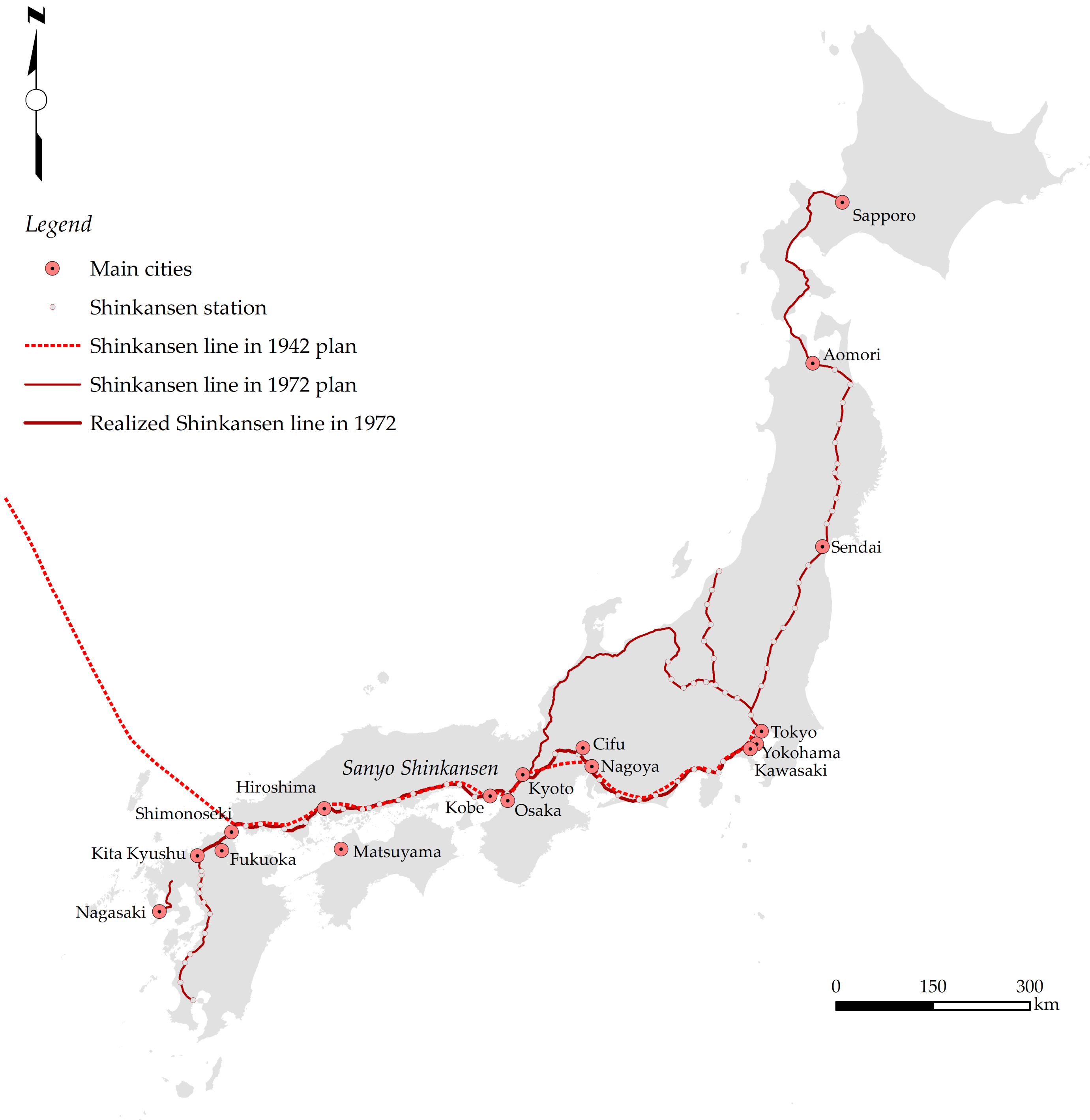

Figure 3: Plans for the Shinkansen network

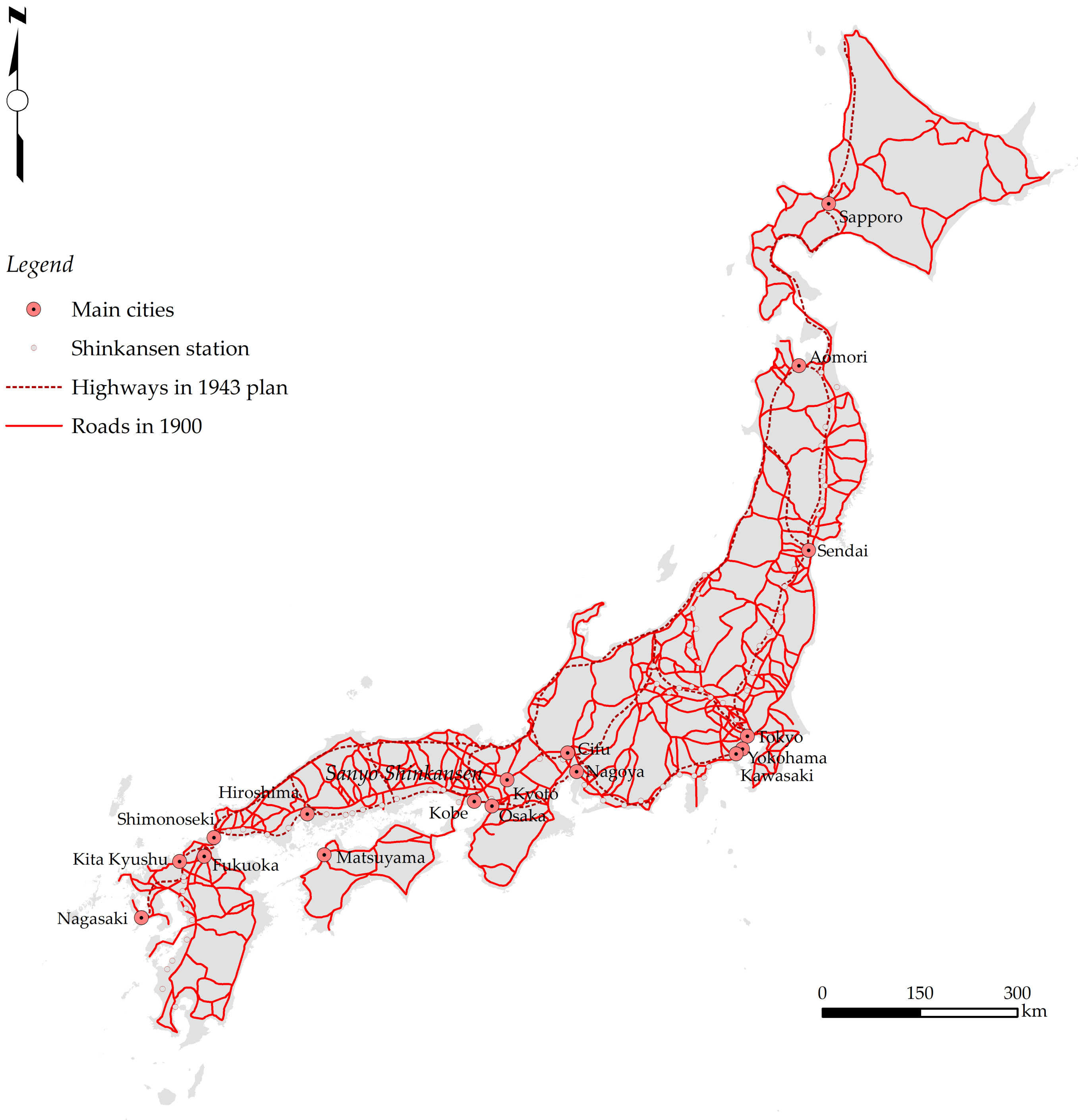

Figure 4: Highway plans and roads in 1900

analysis). We therefore look at the current distribution of Shinkansen stations along Shinkansen lines. It appears that the median network distance to the nearest station is exactly 25 km. Hence, to determine where the Shinkansen stations could be in the Tanaka plan, we evenly distribute stations with 25 km intervals in between the main cities indicated in this plan. We remove all predicted stations that are within 25 km of a main city and/or are not on land.

As a second instrument, we follow Faber (2014) in constructing a least-cost-path spanning tree (LCPST) network. We use elevation maps, where it is increasingly costly to construct railway lines on high altitudes (or through water).[14] Using Dijkstra's algorithm, we construct two LCPST networks: the first one links all cities in the most efficient manner and the other one links Tokyo to all Japanese cities in the least costly way. We integrate those networks and remove overlapping links. Again, we have to determine the station locations, which we do by evenly distributing stations within 25 km intervals on the predicted lines. Figure 4 shows that the LCPST network overlaps pretty much with the Tanaka plan, which we find reassuring. We report regression results in Table C.4.

TABLE C.4 — ALTERNATIVE INSTRUMENTS; SECOND-STAGE RESULTS
(Dependent variable: the log of the change in the employment share)

	(1) 2SLS	(2) 2SLS	(3) 2SLS
Shinkansen station in 2014	**-0.9074**	**-0.8723**	**-1.7523**
	(**0.7884**)	(**0.7836**)	(**2.2543**)
Highway in 2014	-0.2022**	-0.2025**	-0.1946**
	(0.0948)	(0.0948)	(0.0976)
Population density in 1957 *(log)*	-0.0676**	-0.0679**	-0.0598
	(0.0339)	(0.0339)	(0.0395)
Geography controls (4)	Yes	Yes	Yes
Number of observations	1,412	1,412	1,412
Kleibergen-Paap F-statistic	5.369	10.67	2.556

Notes: We exclude municipalities that are centers of MEAs or McEAs. **Bold** indicates instrumented. We use the Tanaka plan instruments in columns (1) and (2); and the LCPST network in columns (1) and (3). Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

The coefficient indicates that a Shinkansen station leads to a change in employment of $\exp(-0.907) - 1 = -60\%$, so the effect is stronger. However, the coefficient is also quite imprecisely estimated, implying that this estimate is not statistically significantly stronger than the baseline estimates (and not significant at conventional levels). A possible explanation for this

[14] The assumptions on how costly it is to construct railway lines at higher altitudes are arbitrary, but our results are very robust to alternative assumptions on the cost scheme. Now, we assume that building a Shinkansen on altitudes exceeding 2000 m or on water bodies is 4 times more expensive, between $1500 - 2000$ m it is 3 times more expensive, between $1000 - 1500$ m, it is twice as expensive, as building on lower grounds.

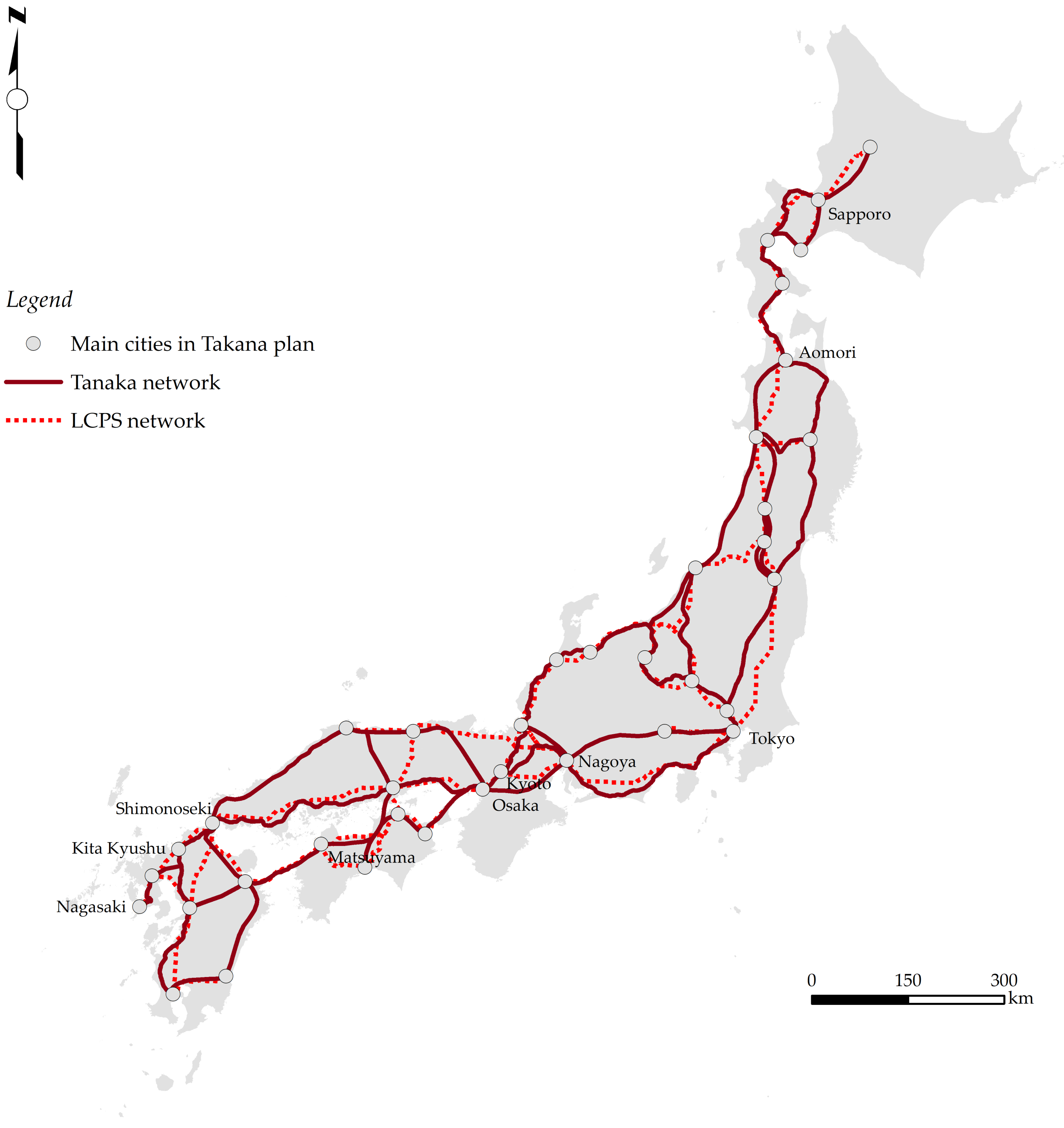

Figure 5: Tanaka plan and the Least-Cost-Path Spanning Tree network

is that the first-stage F-statistic is below the rule-of-thumb value of 10, so that we may have a weak instrument problem. However, if we use estimators that are less susceptible to the weak instrument problem (such as Limited Information Likelihood), the results are nearly identical. When looking at the first-stage results in Table C.5, we observe that having a predicted station in the Tanaka plan has a reasonably strong impact on the probability to have a station in 2014. The coefficient indicates that a station in the Tanaka plan raises the probability to have a station in 2014 by 8.4 percentage points. Over and above the Tanaka plan, the LCPST network does not have much predictive power, which is not surprising given that the Tanaka plan and LCPST network strongly overlap. However, when we only consider the LCPST network in column (3) of Table 5, it appears that a predicted station on the LCPST network raises the probability to have a station in 2014 by 3.7 percentage points.

TABLE C.5 — ALTERNATIVE INSTRUMENTS; FIRST-STAGE RESULTS
(Dependent variable: Shinkansen station or line in 2014)

	(1) OLS	(2) OLS	(3) OLS
Shinkansen station in Tanaka plan	0.0844***	0.0870***	
	(0.0260)	(0.0266)	
Shinkansen station on LCPST network	0.0103		0.0372
	(0.0206)		(0.0233)
Highway in 2014	0.0043	0.0046	0.0074
	(0.0093)	(0.0091)	(0.0096)
Population density in 1957 *(log)*	0.0083	0.0082	0.0094
	(0.0057)	(0.0057)	(0.0058)
Geography controls (4)	Yes	Yes	Yes
Number of observations	1,415	1,415	1,415
R^2	0.0470	0.0467	0.0165

Notes: We exclude municipalities that are centers of MEAs or McEAs. Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Going back to Table C.4, we find that if we use either the Tanaka plan or the LCPST network in columns (2) and (3), respectively, we find very similar results. However, in column (3), because of weak instruments, the coefficient becomes very imprecise and unreasonably strong.

In sum, we find results that are very much in line with the baseline results presented earlier. The coefficients estimated suggest somewhat stronger effects, although the fact that the instruments are less strong also implies that those estimates are not statistically significantly stronger than the preferred estimates.

C.4 Robustness for commuting zones and minryoku areas

In Table C.6, we repeat the baseline instrumental variable regressions for different levels of geographical aggregation. In so doing, we address the concern that there may be commuting occurring between municipalities, so that we mismeasure local labor supply. First, Adachi *et al.* (2020) construct commuting zones using data on commuting flows and a hierarchical agglomerative clustering method. We use their 2015 definition of commuting zones, implying that there are 265 spatial units. In our final sample based on mainland Japan, we end up with 221 spatial units. A downside of this approach is that we cannot exclude central municipalities because we would end up with too small a dataset.

TABLE C.6 — ROBUSTNESS FOR COMMUTING ZONES AND MINRYOKU AREAS
(Dependent variable: the log of the change in the employment share)

	Commuting zones			Minryoku areas		
	(1)	(2)	(3)	(4)	(5)	(6)
	2SLS	2SLS	2SLS	2SLS	2SLS	2SLS
Shinkansen station in 2014	**-0.1517**	**-0.1276**	**-0.3816****	**-0.4037****	**-0.4163****	**-0.1573**
	(**0.0925**)	(**0.0942**)	(**0.1549**)	(**0.1720**)	(**0.1799**)	(**0.3863**)
Highway in 2014	-0.0559	-0.0583	-0.2801	-0.5239	-0.5234	-0.7505***
	(0.4899)	(0.4907)	(0.2690)	(0.3389)	(0.3389)	(0.2282)
Population density in 1957 *(log)*	0.1828***	0.1813***	0.2473**	-0.1076**	-0.1074**	-0.2430***
	(0.0599)	(0.0598)	(0.1173)	(0.0452)	(0.0451)	(0.0497)
Geography controls (4)	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	220	220	96	433	433	304
Kleibergen-Paap F-statistic	460.4	178.6	105.1	35.60	39.42	6.220

Notes: Commuting zones are based on the 2015 definition from Adachi *et al.* (2020). We exclude minryoku-areas that contain municipalities that are centers of MEAs or McEAs. Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. **Bold** indicates instrumented. In columns (1) and (2) we instrument a Shinkansen station in 2014 with dummy indicating whether the minryoku-area has a Shinkansen station planned in the 1942 plan and whether it has one planned in the 1972 plan. In columns (2) and (5) we only instrument with the 1972 plan, while in columns (3) and (6) we only include prefectures on East-Honshu and instrument with the 1942 plan. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In the first three columns of Table C.6, we still find a negative effect ranging from 11 to 32%, which is in the same ballpark as the baseline estimates using municipalities. However, because we cannot exclude central municipalities when aggregating our data up to the commuting zone level, we also consider so-called *minryoku areas*. Minryoku areas are constructed by the company *Asahi Shimbun Shuppan* until 2016 and are also considered as commuting areas. Because minryoku areas are smaller than the commuting zones compiled by Adachi *et al.* (2020), we can exclude minryoku-areas that have a municipality that belongs to a central MEA or McEA city. We end up with a sample of 433 minryoku-areas. The results are very similar with point estimates that are close to our baseline estimates.

In sum, it seems fair to conclude that, when aggregating data within commuting zones or minryoku-areas, a Shinkansen station still has a strong negative effect of on employment growth.

C.5 Addition robustness checks

We perform a set of additional robustness checks in Table C.7. In column (1) we include McEA centers. This increases the number of observations, while McEAs are unlikely to be a reason for a Shinkansen line to have been constructed. This increases the number of included municipalities by 120 (8%), but leaves the results pretty much unaffected. Column (2) excludes the three largest metropolitan areas: Tokyo, Osaka and Nagoya. In this way, we addresses the concern that our results are driven by these large metropolitan areas where commuting between municipalities is ubiquitous. The point estimate is still sizably negative, but imprecisely estimated due to a much lower number of observations. In column (3), we keep municipalities that have a Shinkansen station *or* are located further than 10km from a station. In this way, we avoid the concern that we just measure local effects of employment moving across the municipal borders. The negative coefficient is almost identical to the baseline estimates.

TABLE C7 — ADDITIONAL ROBUSTNESS CHECKS
(Dependent variable: the log of the change in the employment share)

	(1)	(2)	(3)	(4)	(5)	(6)
	2SLS	2SLS	2SLS	2SLS	2SLS	
	Including McEA centers	Excluding top 3 MEAs	Areas >10km station	Controlling for the change in population shares		Exclude outliers
Shinkansen station in 2014	**-0.2650****	**-0.2157**	**-0.3382***	**-0.2737****	**-0.2645****	**-0.3921*****
	(**0.1313**)	(**0.1816**)	(**0.1745**)	(**0.1131**)	(**0.1135**)	(**1768**)
Δ Population share *(log)*				1.0333***	**1.1295*****	
				(0.0225)	(**0.0942**)	
Highway in 2014	-0.2203**	-0.0591	-0.1650	-0.1034**	-0.1028**	-0.2000***
	(0.0935)	(0.0976)	(0.1096)	(0.0501)	(0.0511)	(0.0810)
Population density in 1957 *(log)*	-0.0724**	-0.0481	-0.0074	-0.2806***	-0.2998***	-0.0300
	(0.0318)	(0.0472)	(0.0452)	(0.0254)	(0.0344)	(0.0274)
Geography controls (4)	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	1,532	1,164	1,147	1,412	1,408	1,384
Kleibergen-Paap F-statistic	48.38	21.51	5005	27.69	19.68	27.70

Notes: Geography controls include the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. **Bold** indicates instrumented. We construct instruments based on the 1942 and 1972 plans. In column (5) we instrument for the change in the population share with the log population density in 900. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In column (4) we hope to address the concern that our results are driven by changes in population (e.g. due to migration). We therefore control for the change over time in population shares. We find an elasticity of about one. However, our main results are not much affected. The opening of a Shinkansen is now associated with a decrease in employment of 24%. To the extent one is concerned that our results are now subject to the issue of reverse causality (because jobs may attract people), we instrument for the *change* in population shares with population density in 900 in column (5). The results are hardly affected.

We want to ensure that our results are not driven by potential mismeasurement and/or outliers in employment changes. In column (6) of Table C7, we therefore exclude the bottom and top 1% values of employment changes, leading to a very similar result.

Note that, while the effect of having a Shinkansen station is robust across specifications, the effect of highways is not. It is often much smaller in magnitude than the original estimates. Hence, we may conclude that the effect of highways on employment is not very robust across specifications, although the point estimate generally suggests a negative effect.

C.6 Bias-corrected estimates

Here we aim to investigate whether our OLS results are subject to omitted variable bias. Oster (2019) argues that what matters is the variance of the added control variables, as well as coefficient movements after inclusion of relevant controls. Oster (2019) then derives a GMM estimator to correct the estimates for omitted variable bias under the assumption that the relationship between the variable of interest and observables is informative on the relationship between the variable of interest and unobservable control variables.

There are two key input parameters that have to be determined. First, there is the maximum R^2 from a hypothetical regression of employment changes on the variable of interest as well as all relevant controls, both observable and unobservable. Following Oster (2019) we multiply the R^2 from an OLS regression with all observable controls by 1.3, implying that $R^2_{\max} = 0.3290$. Second, a parameter must be chosen that determines the relative degree of selection on observed and unobserved variables, which we denote by ψ. Altonji *et al.* (2005) and Oster (2019) show that $\psi = 1$ is a reasonable (upper-bound) value. We show bias-corrected estimates for different values of ψ in Figure 6.

We find that bias-corrected estimates are only slightly stronger than the baseline OLS results. For $\psi = 1$, $\hat{\beta} = -0.2719$, which is close to the OLS estimate. Hence, these analyses suggest that, to the extent our observable controls are informative on the unobservables, omitted variable bias is not a big issue.

C.7 Non-parametric regressions

In this Appendix section, we aim to verify an implication of Proposition 2 that the effect of a Shinkansen is likely more positive if the initial size of the region is larger. We therefore let the parameters of the regression model be dependent on $\theta_{i\underline{t}}$. We estimate the following regression:

$$\log \frac{\lambda_{i\bar{t}}}{\lambda_{i\underline{t}}} = f_{\theta_{1957}}\left(S_{i\bar{t}}, h_{i\bar{t}}, \frac{L_{i\underline{t}}}{A_i}, G_i\right) + \varepsilon_{i\underline{t}\bar{t}}, \tag{20}$$

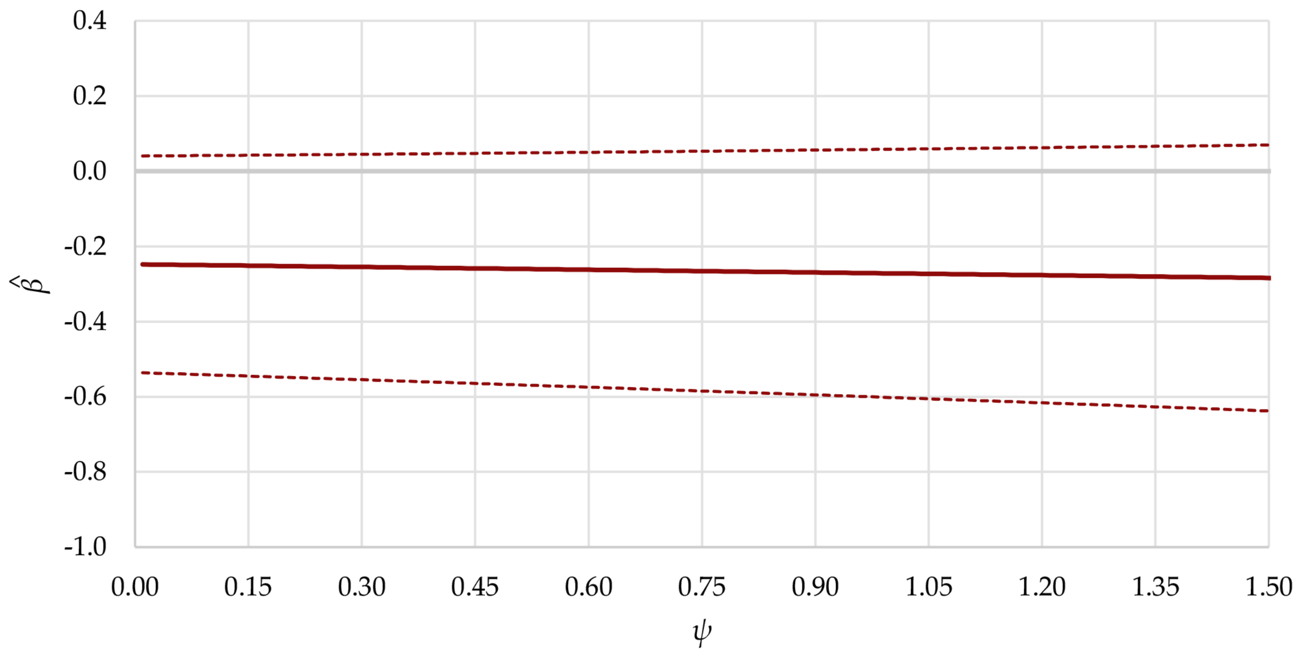

Figure 6: Bias-adjusted estimates

where $f_{\theta_{1957}}(\cdot)$ is an unspecified function. We then use local linear estimation techniques to determine $f_{\theta_{1957}}(\cdot)$. That is, for each observation i we estimate a weighted regression where the weights are defined based on a Gaussian weighting function:

$$w_{ij} = \frac{1}{bs_\theta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta_{i\underline{t}} - \theta_{j\underline{t}}}{bs_\theta}\right)^2}, \tag{21}$$

where w_{ij} is a weight of observation j in a regression of observation i, s_θ is the sample standard deviation of $\theta_{i\underline{t}}$, and b denotes the bandwidth. The bandwidth defines the 'smoothness' of the function to be estimated. When $b \rightarrow \infty$, we arrive at identical estimates as the OLS results, while $b = 0$ would imply that we estimate a separate regression with one observation for each individual observation i. After eye-balling the results, we choose $b = 1$, but our results are not very sensitive to the bandwidth.

An issue with (20) is that $S_{i\bar{t}}$ is endogenous. Following Blundell and Powell (2003), among others, we use a control function approach, where the first-stage errors of a regression of $S_{i\bar{t}}$ on the instruments and controls are inserted in the second stage. However, $S_{i\bar{t}}$ is not continuously distributed, which is an assumption necessary to identify the effects of interest. Tchetgen Tchetgen (2014) however shows that the first stage then can be estimated by a probit model. The residual is then the treatment variable minus the propensity score, which should be inserted as a control variable in the second stage.

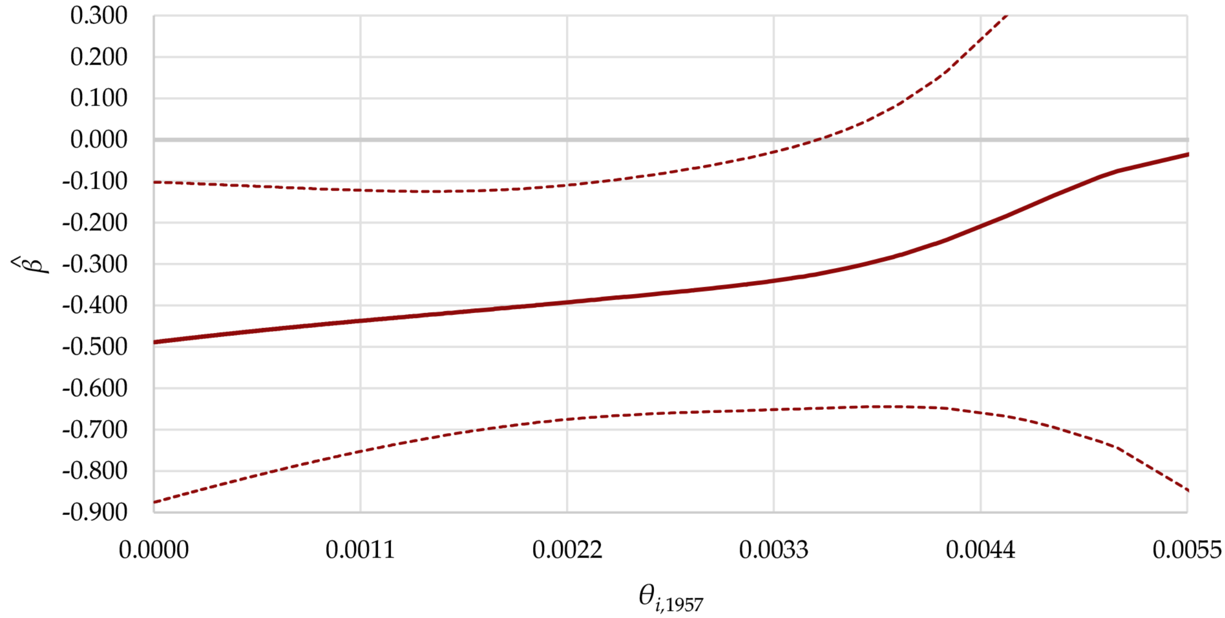

Figure 7: THE EFFECT OF BEING CONNECTED FOR SMALL AND LARGER MUNICIPALITIES

We report results in Figure 7. The dashed lines denote 95% confidence bands. In line with the predictions of Proposition 2, we find evidence that the effect becomes smaller and approaches zero for large municipalities. However, we caution that the confidence bands are wide, so that we cannot draw firm conclusions.

C.8 Spatial heterogeneity

One may wonder whether we may say something on which locations benefit from the opening of a Shinkansen station. Our model does not make strong predictions on this. Still, we hope to shed some light on the issue by classifying locations as being *(i)* a central municipality in an MEA, *(ii)* a suburban municipality in an MEA, *(iii)* a municipality in an McEA, or *(iv)* a 'remote' area. To this end, we estimate the following specification:

$$\log \frac{\lambda_{i\bar{t}}}{\lambda_{i\underline{t}}} = \alpha + \sum_{C=1}^{4} \beta_C (1 - S_{i\bar{t}}) + \gamma h_{i\bar{t}} + \delta \log \frac{L_{i\underline{t}}}{A_i} + \zeta G_i + \varepsilon_{i\underline{t}\bar{t}}. \tag{22}$$

Hence, instead of investigating the impact of a Shinkansen station, we compare the impact of having not received a Shinkansen station. The comparison group consists of municipalities that have received a Shinkansen station. We now have four variables of interest that are potentially endogenous. However we cannot use instrumental variables, because we do not have four instruments. Although we will also obtain Oster's bias-adjusted estimates we will consider our estimates as suggestive. The results are reported in Table C.8.

Table C.8 — Spatial heterogeneity
(Dependent variable: the log of the change in the employment share)

	OLS		Bias-adjusted estimates	
	(1)	(2)	(3)	(4)
	OLS	OLS	GMM	GMM
(1−Shinkansen station in 2014)×	-0.4055***	-0.2962***	-0.2090**	-0.1760
Remote area	(0.0634)	(0.0736)	(0.1047)	(0.1192)
(1−Shinkansen station in 2014)×	-0.1825***	-0.1393*	-0.1091	-0.0985
In McEA	(0.0671)	(0.0729)	(0.0886)	(0.0957)
(1−Shinkansen station in 2014)×	0.5635***	0.4114***	0.3080***	0.2722***
In MEA, suburban	(0.0654)	(0.0670)	(0.0823)	(0.0915)
(1−Shinkansen station in 2014)×	0.1479	0.2114**	0.2527***	0.2666***
In MEA, central area	(0.0911)	(0.0877)	(0.0973)	(0.1032)
Control variables (6)	No	Yes	Yes	Yes
Number of observations	1,412	1,412	1,412	1,412
R^2	0.2129	0.3271		
$R^2_{\max}$			0.4252	0.4252
ψ			0.75	1.00

Notes: Controls include highway access in 2014, the log of population density in 1957, the share of developable land, precipitation in mm per km^2, the mean temperature in January and July. Robust standard errors in columns (1) and (2); and bootstrapped standard errors in columns (3)-(5) (250 replications), are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

In column (1) we show that remote areas and municipalities in McEAs lose employment relative to municipalities that are connected to the Shinkansen network. However, suburban municipalities have faster employment growth compared to municipalities with a Shinkansen station. In column (2) we add the familiar controls related to highway access, population density in 1957 and geography. The results do not qualitatively change much.

Columns (3) and (4) represent estimates adjusted for omitted variable bias, which are similar to the OLS estimates. The estimate in column (4) (i.e. with the relative degree of selection on observed and unobserved variables set to 1) suggests that remote areas lose $\exp(-0.1760) - 1 = -16\%$ in terms of employment relative to Shinkansen station areas. This effect is only 9.4% for municipalities in McEAs. The results that really stand out are the coefficients related to suburban and central areas. We find that those areas see an employment growth of about 30% compared to municipalities that have received a station. These results are generally in line with Hayakawa *et al.* (2020) who, using a spatial quantitative general equilibrium model, find that central and suburban areas generally have benefited from the Shinkansen network.